Quelques résultats en statistiques des grandes dimensions

Marc-Antoine Giuliani

UNIVERSITÉ PARIS DIDEROT (PARIS 7)

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Laboratoire de Modèles Aléatoires et de Probabilités - CNRS UMR 7599

THÈSE DE DOCTORAT

Discipline : Mathématiques Appliquées

Présentée par

Marc-Antoine Giuliani

QUELQUES RÉSULTATS EN STATISTIQUES DES GRANDES DIMENSIONS

Sous la direction de **Dominique PICARD**

Rapporteurs :

M. **Pierre**	**ALQUIER**	ENSAE
M. **Sylvain**	**SARDY**	Université de Genève

Soutenue publiquement le **24 mai 2016** devant le jury composé de :

M. **Pierre**	**ALQUIER**	ENSAE	Rapporteur
M. **Stéphane**	**BOUCHERON**	Université Paris Diderot	Examinateur
M. **Erwan**	**LE PENNEC**	Ecole Polytechnique	Examinateur
Mme **Dominique**	**PICARD**	Université Paris Diderot	Directrice
M. **Sylvain**	**SARDY**	Université de Genève	Rapporteur

Remerciements.

Mes tout premiers remerciements vont à ma directrice Dominique Picard. Merci Dominique d'avoir encadré ce travail, de m'avoir accompagné tout au long de ces quatre ans, je suis bien conscient de n'avoir pas été un élève facile ! Mais ta gentillesse, ton savoir et ton investissement constant m'ont permis de mener ce travail au bout. Merci aussi de m'avoir donné l'opportunité très enrichissante d'assurer un de tes TD pendant mon année d'ATER.

Je remercie mes rapporteurs Pierre Alquier et Sylvain Sardy pour le travail qu'ils ont eu le courage d'accomplir.

Je remercie Stéphane Boucheron et Erwan Le Pennec d'avoir bien voulu faire partie de mon jury.

Il est d'usage de remercier tous ceux qui nous ont apporté de la connaissance, de l'aide et de la joie de manière plus ou moins alphabétique, permettez moi de les remercier par ordre chronologique. Tout d'abord (et à nouveau) merci Dominique et Pierre de m'avoir fait découvrir le sujet passionnant de la statistique mathématique, vos enseignements pendant mon année de M1 ont été déterminants. Merci Pierre pour tes encouragements et ta bienveillance (et tes tampons smileys !). Merci Monsieur Kerkyacharian et Erwan de m'avoir fait découvrir les ondelettes et plus largement le sujet de l'estimation non-paramètrique. Merci pour votre enseignement riche et exigeant. Merci Erwan de m'avoir encouragé à me lancer dans cette aventure de la thèse. Merci Stéphane pour tes enseignements, ta gentillesse permanente, ton bureau était toujours ouvert (et ton goût du code rassérénant). Merci Mathilde de m'avoir appris R et plus largement de m'avoir fait réaliser que notre discipline n'a de sens qu'appliquée. Merci Noufel pour ta bonne humeur constante et les nombreux cafés pris ensemble. Merci Aurélie de m'avoir accompagné dans mon activité d'enseignement. Merci Maud et Lorick d'avoir sû créer une merveilleuse ambiance alors que nous commencions tous les trois l'aventure du doctorat. Merci Thomas, ton sens de l'humour et ton goût des maths ont été un soutien précieux. Merci Guillaume d'avoir toujours aimé résoudre des problèmes de maths, petits ou grands ! Enfin merci Anna pour ton soutien constant et ton amitié.

Je tiens à remercier mes parents, mon père qui a été mon premier professeur de mathématiques et ma mère qui a toujours été un soutien indéfectible. Un grand merci à ma soeur Mathilde de m'avoir toujours encouragé, et d'avoir été là dans les moments difficiles. A mes amis Anne, Brice, Charles, Gaspard, Jean-Paul, Simon, Stéphane pour tous nos bols d'air du week-end. J'ai de plus une pensée particulière pour ma grand-mère qui aurait aimé être là aujourd'hui.

Enfin et tout particulièrement merci à toi Annabelle, c'est autant ta thèse que la mienne (mais j'ai fait toutes les démarches administratives pour une fois !). Tu as été là à chaque pas et sans toi rien n'aurait été possible. Merci pour ces années passées et pour toutes celles à venir.

A tous ceux que j'ai côtoyés et qui m'ont soutenu à un moment ou un autre de ces quatre années, merci et qu'ils m'excusent de ne pouvoir tous les nommer.

Table des matières

Chapitre 1

Introduction

Sommaire

1 Le modèle linéaire

1.1 Motivation du modèle

Le modèle linéaire est l'un des objets centraux de la statistique mathématique. Il est le parfait exemple d'une construction inspirée d'un problème concret et autour de laquelle

une riche théorie s'est mise en place : étant donné une variable d'intérêt, y, peut-on mesurer l'influence qu'ont sur elle une famille de p covariables, $x_1, \ldots, x_p$?

Bien entendu, pour que la chose soit possible, il est nécessaire de supposer que y et les covariables x_i sont liées. Le modèle linéaire est une façon de spécifier ce lien en supposant que :

$$y = \alpha_1^* x_1 + \cdots + \alpha_p^* x_p + \varepsilon, \tag{1.1}$$

où les α_i^* sont des scalaires et où ε est un terme de bruit, que l'on modélisera comme une variable aléatoire réelle d'espérance nulle.

Le statisticien dispose d'un échantillon d'apprentissage de n observations de la variable y et des covariables x_i. Il est pratique de noter $\mathbf{y} \in \mathbb{R}^n$ le vecteur d'observations de la variable y dans cet échantillon d'apprentissage, de même on note $\mathbf{x}_i \in \mathbb{R}^n$ le vecteur d'observations de la covariable x_i. Il est alors naturel de regrouper les observations des covariables dans une matrice (dite de design) :

$$\mathbf{X} = \begin{pmatrix} \vdots & \vdots \\ x_{i1} & x_{ip} \\ \vdots & \vdots \end{pmatrix} = [\mathbf{x}_1 \ldots \mathbf{x}_p],$$

dont les colonnes sont les vecteurs $\mathbf{x}_i$. Alors la relation eq. (1.1) se traduit en un système d'équations dans l'échantillon d'apprentissage :

$$\mathbf{y} = \mathbf{X}\,\alpha^* + \boldsymbol{\varepsilon}, \tag{1.2}$$

où $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est un vecteur dont chaque composante représente le bruit associé à chaque observation de la variable y, et $\alpha^* \in \mathbb{R}^p$ est le vecteur des paramètres. Sur la figure 1.1 on représente une telle relation sur un jeu de données simulées.

On peut maintenant reformuler la question initiale plus précisément : **étant donné un échantillon d'apprentissage de taille n, comment estimer le mieux possible la valeur des p paramètres α_i^* si l'on suppose la relation 1.1 entre la variable d'intérêt y et les covariables $x_1, \ldots, x_p$?**

1.2 Erreur de prédiction, cadre minimax

Soit $\hat{\alpha}$ un estimateur du vecteur de paramètres α^*, c'est-à-dire que $\hat{\alpha}$ est un vecteur de $\mathbb{R}^p$ construit uniquement à partir de l'observation de $\mathbf{y}$ et de $\mathbf{X}$ (de façon mesurable). Afin de répondre à la question précédente on doit se doter d'un critère d'erreur pour mesurer la qualité de $\hat{\alpha}$ et être à même de comparer si possible deux estimateurs entre eux.

Un critère raisonnable pour mesurer la qualité d'un estimateur est de regarder à quel point il est capable de bien prédire l'espérance de la variable d'intérêt sachant la valeur des

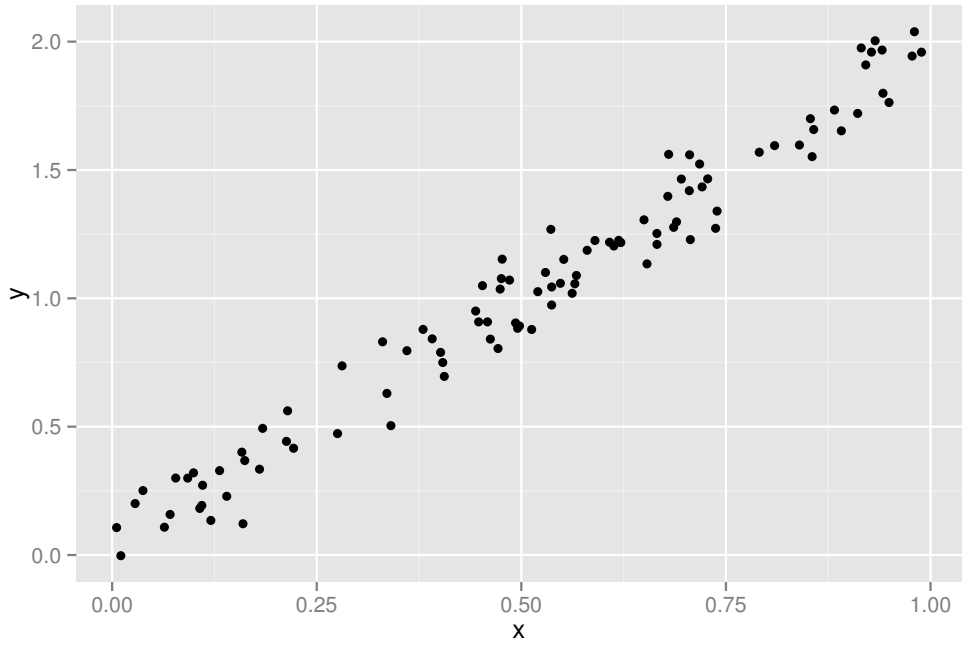

FIGURE 1.1. Jeu de données simulées : $\mathbf{y} = 2\mathbf{x} + \boldsymbol{\varepsilon}$.

covariables sur les données de l'échantillon d'apprentissage. L'erreur quadratique moyenne (Mean Squared Error), $\frac{1}{n}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}\|_2^2$, est justement une façon naturelle de quantifier l'écart entre le vecteur d'intérêt $\mathbf{X}\alpha^*$ et sa prédiction $\mathbf{X}\hat{\alpha}$. Mais cette quantité est aléatoire, on cherchera donc a borner son espérance, qu'on qualifiera de risque de prédiction :

$$R(\alpha^*, \hat{\alpha}) = \mathbb{E}\Big[\frac{1}{n}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}\|_2^2\Big]. \tag{1.3}$$

On parle de risque de prédiction car si le design $\mathbf{X}$ est représentatif des valeurs prises par les covariables, cette quantité reflète bien la capacité de l'estimateur $\hat{\alpha}$ à fournir de bonnes prédictions sur des observations futures.

Il est bien entendu sans intérêt de parler d'optimalité d'un estimateur en un point $\alpha^* \in \mathbb{R}^p$. En effet l'estimateur déterministe $\hat{\alpha} = \alpha^*$ est toujours optimal au point α^*, et pourtant n'est pas du tout efficace en tout autre $\alpha \in \mathbb{R}^p$ assez éloigné de α^*. Pour pallier cette difficulté on introduit la notion d'optimalité au sens minimax, qui caractérise à quelle vitesse α^* peut être estimé uniformément sur un certain sous-ensemble Λ de $\mathbb{R}^p$.

Definition 1.1. *On dira qu'un estimateur $\hat{\alpha}$ est optimal au sens minimax sur Λ si :*

$$R(\alpha^*, \hat{\alpha}) = \mathbb{E}\Big[\frac{1}{n}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}\|_2^2\Big] \leq C\psi_{n,p},$$

pour une certaine suite $(\psi_{n,p})$, et une constante $C > 0$, et s'il existe une constante $C' > 0$ telle que :

$$\inf_{\hat{\beta}} \sup_{\alpha^* \in \Lambda} \mathbb{E}\Big[\psi_{n,p}^{-1}\frac{1}{n}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\beta}\|_2^2\Big] \geq C',$$

où l'infinimum est pris sur tous les estimateurs $\hat{\beta}$ possibles. De plus $\psi_{n,p}$ est appelée vitesse d'estimation minimax sur Λ.

Un estimateur est donc minimax s'il est celui dont la pire erreur sur Λ est la moins grande.

2 Estimation par projection

Nous commençons par introduire l'estimateur des moindres carrés dans le modèle linéaire précédemment décrit. Nous essayons alors d'expliquer en quoi, bien que tout à fait raisonnable pour les situations où l'on dispose de beaucoup plus d'observations qu'il n'y a de covariables, cet estimateur n'est pas adapté aux problématiques contemporaines de grandes dimensions, où le nombre de covariables est grand devant le nombre d'observations. La section suivante tente de remédier à ce problème en postulant la parcimonie du vecteur α^* et en adaptant la procédure d'estimation.

2.1 Estimateur des moindres carrés

On cherche à construire ici un estimateur qui minimise le risque de prédiction :

$$R(\alpha^*, \hat{\alpha}) = \mathbb{E}\Big[\frac{1}{n}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}\|_2^2\Big].$$

Le vecteur $\mathbf{X}\alpha^*$ n'étant pas directement observable, une idée naturelle est alors de le remplacer par le vecteur observé $\mathbf{y} = \mathbf{X}\alpha^* + \boldsymbol{\varepsilon}$. On cherche donc un estimateur qui soit solution du problème de minimisation :

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\alpha\|_2^2. \tag{1.4}$$

Toute solution du problème 1.4 est appelée estimateur des moindres carrés (puisqu'il minimise le carré d'une norme euclidienne), et sera notée $\hat{\alpha}^{\mathrm{LS}}$. Le théorème de projection dans les espaces de Hilbert garantit l'existence de ces estimateurs. De plus il implique que tout estimateur $\hat{\alpha}^{\mathrm{LS}}$ vérifie la relation :

$$\mathbf{X}\hat{\alpha}^{\mathrm{LS}} = P_{V_{\mathbf{X}}}[\mathbf{y}], \tag{1.5}$$

où $V_{\mathbf{X}}$ est l'espace vectoriel image du design $\mathbf{X}$, et $P_{V_{\mathbf{X}}}$ est le projecteur orthogonal sur $V_{\mathbf{X}}$.

Un estimateur des moindres carrés jouit de nombreuses bonnes propriétés, il est par exemple optimal dans la classe des estimateurs linéaires non biaisés de α^* lorsque $\mathbf{X}$ est une injection (théorème de Gauss-Markov [85]). A l'inverse, le point de départ des méthodes que nous étudierons ensuite vient de son incapacité à s'adapter aux problèmatiques dites de "grandes dimensions". Le théorème suivant fournit une borne sur son erreur de prédiction qui met en lumière ce phénomène.

Théorème 1.2. *Supposons que la relation 1.1 soit vraie, et supposons de plus que le terme de bruit est gaussien de variance σ^2, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Alors l'estimateur des moindres carrés vérifie :*

$$R(\alpha^*, \hat{\alpha}^{LS}) \lesssim \sigma^2 \frac{r}{n}, \tag{1.6}$$

où r est le rang de ${}^t\mathbf{X}\mathbf{X}$.

Ce résultat a deux interprétations très importantes en fonction de la situation qu'on considère.

Le cas "classique" : pendant longtemps l'étude du modèle linéaire reposait implicitement sur l'idée que le nombre de covariables utilisées était fixe, alors qu'il était relativement aisé d'acquérir plus d'observations. C'est-à-dire que le nombre d'observations pouvait facilement être rendu plus grand que le nombre de covariables, qui n'augmentait pas avec l'acquisition de nouvelles observations. Sous ces conditions, c'est-à-dire si le nombre d'observations est grand devant le nombre, fixe, de covariables alors le théorème 1.2 garantit que le risque de prédiction de l'estimateur des moindres carrés décroit comme $\frac{1}{n}$. Une riche littérature fait l'état des connaissances accumulées dans ce cas, on pourra se référer par exemple à [85] ou à [97] (pour voir la théorie développée dans un espace euclidien général).

Le cas "grandes dimensions" : de plus en plus aujourd'hui, les jeux de données auxquels le statisticien est confronté ne rentrent plus dans le cadre "classique" décrit précédemment. En effet, de nombreux domaines acquièrent des données où le nombre de covariables est grand devant le nombre d'observations. En particulier la génomique où la technologie des puces à ADN permet l'acquisition, pour chaque observation, des niveaux d'expressions d'un grand nombre de gènes. Chaque observation restant plutôt coûteuse (ou la population étudiée étant très restreinte, comme dans le cas de maladies génétiques rares) les données obtenues ne rentrent plus dans le cadre "classique", le nombre de covariables y étant beaucoup plus grand que le nombre d'observations. Le "text mining" est un autre domaine où chaque acquisition d'une nouvelle observation, c'est-à-dire d'un nouveau texte, s'accompagne d'une augmentation du nombre de covariables. En effet dans le modèle dit de "bag of words", chaque texte d'un corpus est une observation alors que les mots dont ils sont constitués forment les covariables. Avec l'acquisition d'un nouveau texte, s'ajoutent aux précédentes covariables, de nouveaux mots non précédemment observés. Dans ce cas le nombre de paramètres p n'est plus fixe, mais croît avec n et peut être beaucoup plus grand que n. Dans ce contexte où p n'est plus fixe, et où potentiellement on peut avoir $p \gg n$, le théorème 1.2 ne garantit plus rien sur la vitesse d'estimation de l'estimateur des moindres carrés.

Afin d'illustrer les difficultés de ce nouveau paradigme, considérons le cas orthogonal où les calculs sont simplifiés. Dans ce cas, on suppose que le nombre de covariables est égal au nombre d'observations n. On suppose de plus que les colonnes du design $\mathbf{X}$ forment une base orthonormale de l'espace $\mathbb{R}^n$. Alors l'estimateur des moindres carrés est unique et se réduit à $\hat{\alpha}^{\mathrm{LS}} = {}^t\mathbf{X}\mathbf{y}$. De plus, sous les hypothèses du théorème 1.2, on peut calculer son risque de prédiction pour tout $\alpha^* \in \mathbb{R}^n$, qui n'est autre que :

$$R(\alpha^*, \hat{\alpha}^{\mathrm{LS}}) = \sigma^2.$$

L'estimateur des moindres carrés ne voit même plus son risque tendre vers 0 lorsque n

tend vers l'infini. De plus on peut prouver que cet estimateur est minimax sur $\mathbb{R}^n$, nous n'avons donc pas d'espoir de construire un meilleur estimateur (au sens minimax) que $\hat{\alpha}^{\mathrm{LS}}$!

Ainsi, si on ne fait aucune hypothèse a priori sur le vecteur α^*, le problème du modèle linéaire en grandes dimensions est une cause perdue. Heureusement, en pratique les vecteurs de paramètres ne vivent pas dans tout $\mathbb{R}^p$ mais plutôt sur une sous-variété de $\mathbb{R}^p$ de dimension intrinsèque bien inférieure à p. En effet si l'on considère l'exemple des bases d'ondelettes, on sait que la plupart des signaux y admettent une représentation qui utlise peu de coefficients, qu'on qualifie de sparse. Même si le signal vit initialement dans $\mathbb{R}^p$, une fois transformé de la sorte, il appartient donc à l'ensemble des signaux n'ayant que $k \ll n$ coefficients non nuls. Si jamais l'on pouvait deviner a priori où se situent ces k coefficients, alors on pourrait directement effectuer la régression linéaire en n'utilisant que ces k covariables et le théorème 1.2 garantirait une vitesse en $\frac{k}{n} \ll 1$. On peut alors espérer que sous une hypothèse de sparsité du vecteur des paramètres, c'est-à-dire sous l'hypothèse que seul un petit nombre des covariables sont en fait nécessaires, on puisse contruire des estimateurs détectant quels paramètres doivent être estimés, et estimant seulement ceux là. On obtiendrait alors de bien meilleures propriétés asymptotiques que l'estimateur des moindres carrés, puisque le nombre total de covariables serait remplacé par la sparsité réelle de α^* dans la vitesse de convergence.

2.2 Estimation parcimonieuse

Cette partie s'inspire très largement de [18]. On introduit maintenant le concept d'estimateur des moindres carrés restreint. Soit C est un convexe fermé de $\mathbb{R}^p$. On peut considérer le problème de minimisation :

$$\hat{\alpha}_C = \arg\min_{\alpha \in C} \|\mathbf{y} - \mathbf{X}\alpha\|_2^2. \tag{1.7}$$

On parlera d'estimateur des moindres carrés restreint (à C) pour toute solution au problème 1.7. Le théorème de projection sur les convexes fermés d'un espace de Hilbert garantit l'existence d'une solution à ce problème. De plus il garantit que toute solution vérifie :

$$\mathbf{X}\hat{\alpha}_C = P_{C_{\mathbf{X}}}[\mathbf{y}], \tag{1.8}$$

où $P_{C_{\mathbf{X}}}$ est le projecteur orthogonal sur l'ensemble $C_{\mathbf{X}}$, image de C par $\mathbf{X}$. De plus si la restriction de $\mathbf{X}$ à C est injective alors cette solution est unique.

Soit $\mathcal{M}$ un sous-ensemble de $\{1, \ldots, p\}$. Une famille particulièrement importante de sous-espaces vectoriels de $\mathbb{R}^p$ sont les espaces :

$$V(\mathcal{M}) = \{\alpha \in \mathbb{R}^p;\, \alpha_i = 0,\, \forall i \notin \mathcal{M}\}.$$

Les estimateurs des moindres carrés restreints à $V(\mathcal{M})$ sont alors les estimateurs des moindres carrés au sens de 1.4 mais n'utilisant que les covariables indéxées par $\mathcal{M}$! On les

notera $\hat{\alpha}(\mathcal{M})$ dans la suite. Alors la relation 1.8 implique que :

$$\mathbb{E}\Big[\mathbf{X}\hat{\alpha}(\mathcal{M})\Big] = \mathbb{E}\Big[P_{V(\mathcal{M})_{\mathbf{X}}}[\mathbf{y}]\Big] = P_{V(\mathcal{M})_{\mathbf{X}}}[\mathbf{X}\alpha^*]. \tag{1.9}$$

On en déduit la décomposition, particulièrement importante, de l'erreur de prédiction comme :

$$\mathbb{E}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}(\mathcal{M})\|_2^2 = \|\mathbf{X}\alpha^* - P_{V(\mathcal{M})_{\mathbf{X}}}[\mathbf{X}\alpha^*]\|_2^2 + \mathbb{E}\Big[\|P_{V(\mathcal{M})_{\mathbf{X}}}[\boldsymbol{\varepsilon}]\|_2^2\Big]. \tag{1.10}$$

Cette décomposition a l'interprétation habituelle d'une décomposition biais-variance : le terme $\|\mathbf{X}\alpha^* - P_{V(\mathcal{M})_{\mathbf{X}}}[\mathbf{X}\alpha^*]\|_2^2$ s'interprète comme le carré d'un biais, plus $\mathcal{M}$ est grand plus cette quantité est petite, alors que $\mathbb{E}\Big[\|P_{V(\mathcal{M})_{\mathbf{X}}}[\boldsymbol{\varepsilon}]\|_2^2\Big]$ s'interprète comme une variance, qui elle, à l'opposé du biais, croît avec le nombre de covariables utilisées.

Plaçons nous sous les hypothèses du théorème 1.2, en considérant que le vecteur de bruit $\boldsymbol{\varepsilon}$ est gaussien, homoscédastique, de covariance $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Alors la décompositon 1.10 dans ce cas particulier peut s'écrire :

$$\mathbb{E}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}(\mathcal{M})\|_2^2 = \|\mathbf{X}\alpha^* - P_{V(\mathcal{M})_{\mathbf{X}}}[\mathbf{X}\alpha^*]\|_2^2 + \sigma^2 \dim\,(V(\mathcal{M})_{\mathbf{X}}). \tag{1.11}$$

Ainsi chaque covariable que l'on décide d'utiliser ajoute σ^2 à la variance de l'estimateur, mais réduit son biais en contrepartie. Par contre, il n'est absolument pas nécessaire que la réduction du biais apportée par l'introduction d'une covariable soit supérieure à l'augmentation de la variance qu'elle induit.

Supposons maintenant que le vecteur que l'on souhaite estimer, α^*, est k-sparse, c'est-à-dire que ses coefficients non nuls sont indexés par un certain sous-ensemble support $\mathcal{M}^* \subset \{1, \ldots, p\}$, tel que $|\mathcal{M}^*| \leq k$. On notera par la suite $\|\alpha^*\|_0$ le cardinal du sous-ensemble d'indices support de α^*. Alors, si $\dim\,(V(\mathcal{M}^*)_{\mathbf{X}}) = |\mathcal{M}^*| \leq k$, l'estimateur des moindres carrés restreint à $\mathcal{M}^*$ vérifie :

$$\frac{1}{n}\mathbb{E}\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}(\mathcal{M}^*)\|_2^2 \leq \sigma^2\,\frac{k}{n} \ll \sigma^2, \tag{1.12}$$

dès que $k \ll n$. Ainsi, si l'on sait a priori que $\|\alpha^*\|_0 \leq k$, on a intérêt à remplacer l'estimateur des moindres carrés par l'estimateur restreint solution du problème de minimisation :

$$\begin{cases} \hat{\alpha}^S = \arg\min_\alpha \|\mathbf{y} - \mathbf{X}\alpha\|_2^2, \\ \text{s. t. } \|\alpha\|_0 \leq k. \end{cases} \tag{1.13}$$

Cet estimateur $\hat{\alpha}^S$ possède alors, lorsque α^* est bien k-sparse, des propriétés de convergence bien supérieures à l'estimateur des moindres carrés non restreint.

Théorème 1.3. *Plaçons nous sous les hypothèses du théorème 1.2. Supposons que α^* soit k-sparse avec $k \leq p/2$. Alors :*

$$R(\alpha^*, \hat{\alpha}^S) \lesssim \sigma^2 \frac{k}{n} \log\Big(\frac{ep}{k}\Big). \tag{1.14}$$

Dans ce théorème, on trouve bien le terme $\frac{k}{n}$ qui est la vitesse qu'on obtiendrait si l'on savait a priori où se situe le support de α^*. On paye en plus un prix lié au fait qu'on ne connait pas le support de α^* mais seulement une borne sur son cardinal avec le facteur multiplicatif $\log\left(\frac{ep}{k}\right)$. Il suffit maintenant pour avoir convergence de vérifier la condition $\frac{\log p}{n} \to 0$ à sparsité fixée !

L'estimateur $\hat{\alpha}^S$ souffre pourtant de deux défauts rédhibitoires :

1. son calcul repose sur la connaissance a priori de la sparsité (ou du moins d'une bonne borne) de α^*. Il est non adaptatif !
2. même si l'on disposait de cette connaissance a priori, il est pratiquement incalculable. En effet il nécessite le calcul d'un nombre exponentiel d'estimateurs des moindres carrés, un pour chaque sous-ensemble de cardinal inférieur à k de $\{1, \dots, p\}$, c'est-à-dire de l'ordre de p^k estimateurs. Dès que p et k sont modérément grands c'est une tâche irréalisable.

Le problème de la regression sparse peut maintenant se formuler de la façon suivante : comment obtenir un estimateur adaptatif $\hat{\alpha}$ calculable (en temps au plus polynomial) avec une efficacité proche de $\hat{\alpha}^S$?

3 Estimation adaptative pour la régression parcimonieuse

On commence par présenter le cas de la regression parcimonieuse (sparse) avec un design orthogonal qui sert de fondement aux développements ultérieurs et pour lequel une théorie complète existe. On présente ensuite les deux stratégies générales pour fournir des estimateurs effectivement calculables, en temps polynomial, au comportement proche de la solution du problème 1.13 dans le cas d'un design général : les estimateurs obtenus par relaxation convexe de la pénalité ℓ_0 et les estimateurs obtenus par des méthodes greedy. Dans toute la suite on supposera le terme de bruit $\boldsymbol{\varepsilon}$ gaussien.

3.1 Design orthogonal et seuillage

Supposons que les colonnes du design $\mathbf{X}$ forment une base orthonormale de l'espace $\mathbb{R}^n$. Supposons de plus dans un premier temps que le bruit $\boldsymbol{\varepsilon}$ est blanc, c'est-à-dire que $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$. Sous l'hypothèse d'orthonormalité du design, l'estimateur des moindres carrés n'est autre que $\hat{\alpha}^{\mathrm{LS}} = {}^t\mathbf{X}\mathbf{y}$ et vérifie :

$$\hat{\alpha}^{\mathrm{LS}} = \alpha^* + \tilde{\boldsymbol{\varepsilon}}, \tag{1.15}$$

où $\tilde{\boldsymbol{\varepsilon}} = {}^t\mathbf{X}\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Cette relation est qualifiée de modèle de suite gaussienne. De plus on peut remarquer que l'erreur quadratique moyenne en prédiction se réduit ici à

$\|\mathbf{X}\alpha^* - \mathbf{X}\hat{\alpha}\|_2^2 = \|\alpha^* - \hat{\alpha}\|_2^2$. Introduisons les formes seuillées de $\hat{\alpha}^{\mathrm{LS}}$ au niveau λ :

$$t_\lambda^h(\hat{\alpha}^{\mathrm{LS}})_i = \begin{cases} \hat{\alpha}_i^{\mathrm{LS}}, \text{ si } |\hat{\alpha}_i^{\mathrm{LS}}| \geq \lambda, \\ 0, \text{ sinon,} \end{cases} \tag{1.16}$$

est l'estimateur avec seuillage "hard", et

$$t_\lambda^s(\hat{\alpha}^{\mathrm{LS}})_i = \begin{cases} \hat{\alpha}_i^{\mathrm{LS}} - \lambda, \text{ si } \hat{\alpha}_i^{\mathrm{LS}} \geq \lambda, \\ 0, \text{ si } |\hat{\alpha}_i^{\mathrm{LS}}| < \lambda, \\ \hat{\alpha}_i^{\mathrm{LS}} + \lambda, \text{ si } \hat{\alpha}_i^{\mathrm{LS}} \leq -\lambda, \end{cases} \tag{1.17}$$

est l'estimateur avec seuillage "soft". Alors, en utilisant le concept d'inégalité oracle, [37] prouve le résultat suivant.

Théorème 1.4. *Sous les hypothèses du théorème 1.3, si $t_\lambda(\hat{\alpha}^{LS})$ est un estimateur seuillé de $\hat{\alpha}^{LS}$ (hard ou soft) au niveau $\lambda = \sigma\sqrt{2\log n}$ on a :*

$$R(\alpha^*, t_\lambda(\hat{\alpha}^{LS})) \lesssim \sigma^2 \log(n)\frac{k}{n}. \tag{1.18}$$

C'est un résultat tout à fait remarquable car on obtient essentiellement la même vitesse que 1.3 sans avoir à incorporer de savoir a priori sur la sparsité de α^* ! En effet on peut remarquer que l'estimateur avec seuillage hard $t_\lambda^h(\hat{\alpha}^{\mathrm{LS}})$ est aussi solution du problème de minimisation :

$$t_\lambda^h(\hat{\alpha}^{\mathrm{LS}}) = \arg\min_\alpha \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + \lambda^2\|\alpha\|_0 \tag{1.19}$$

qui est la forme lagrangienne de 1.13. C'est-à-dire que pour tout k dans 1.13, il existe un λ tel que 1.19 soit équivalent. Mais le théorème 1.4 fournit une stratégie de seuillage universelle indépendante de la sparsité réelle de α^*, le seuil ne dépendant que du niveau de bruit et de la dimension n, l'estimateur seuillé est adaptatif. De plus il est aisément calculable, le problème 1.13 étant explicitement résoluble sous l'hypothèse d'orthogonalité du design.

En fait le résultat prouvé dans [37] est beaucoup plus fort et permet de prouver la minimaxité de l'estimateur seuillé sous bien d'autres conditions que la sparsité de α^*, en particulier lorsque α^* appartient à une boule de ℓ_q pour $0 \leq q \leq 1$. La théorie du seuillage a été essentiellement développée en vue d'applications à la statistique non paramétrique et une littérature très riche autour de la question de l'estimation non linéaire [33], [32], dans les bases d'ondelettes [28], [70] existe. On pourra se référer à [39], [41], [55], [40] pour ne citer que quelques exemples.

Enfin, la théorie a été ensuite étendue au cas d'un bruit non blanc, c'est-à-dire lorsque le bruit gaussien admet une matrice de covariance non triviale, dans [59], [56], [64], en incorporant l'hétéroscédasticité du bruit dans la stratégie de seuillage.

3.2 La méthode de relaxation convexe

Reconsidérons un instant les résultats de la section précédente. L'estimateur par seuillage soft , $t^s_\lambda(\hat{\alpha}^{\mathrm{LS}})$, qui est adaptatif et minimax sur la classe des α^* sparse, peut se décrire comme la solution du problème de minimisation :

$$t^s_\lambda(\hat{\alpha}^{\mathrm{LS}}) = \arg\min_\alpha \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + 2\lambda\|\alpha\|_1. \tag{1.20}$$

Ici la pénalité ℓ_0 du seuillage hard est relaxée en une pénalité ℓ_1 convexe sans détériorer les bonnes propriétés de l'estimateur. De plus ce problème étant convexe, il est résoluble en temps polynomial quel que soit le design $\mathbf{X}$ (en effet l'optimisation convexe est un sujet important auxquel une grande littérature est consacrée, et où beaucoup d'algorithmes efficaces on été développés [11]). Il est donc naturel pour un design quelconque de chercher l'estimateur solution de :

$$\hat{\alpha} = \arg\min_\alpha \|\mathbf{y} - \mathbf{X}\alpha\|_2^2 + \lambda\|\alpha\|_1. \tag{1.21}$$

Cet estimateur, est qualifié d'estimateur Lasso [89]. De nombreux algorithmes spécifiques ont été développés pour resoudre ce problème de minimisation. On pourra citer en particulier la méthode d'homotopie [44], ou encore plus efficace en pratique la descente de coordonnées [48].

Le problème de cet estimateur réside dans le fait qu'il est a priori adapté à la norme ℓ_1 de α^* et non pas à sa sparsité. La magie de cet estimateur réside elle dans le fait que, si l'on ajoute certaines conditions sur le design, alors il est tout aussi efficace, pour un bon choix de λ, que l'estimateur par seuillage dans le cas d'un design orthogonal. Il existe de nombreux types de conditions exigées sur la matrice de design $\mathbf{X}$, mais tous se résument à réclamer qu'elle ne présente pas de corrélations trop fortes, qu'elle ne soit pas trop éloignée d'une matrice orthogonale. Détaillons en quelques unes :

— il y a les conditions dites de cohérence. La cohérence d'un design est définie comme le plus grand (en valeur absolue) terme extra-diagonal de la matrice de Gram ${}^t\mathbf{X}\mathbf{X}$ (convenablement normalisée). Si l'on suppose que les colonnes de $\mathbf{X}$ sont de norme euclidienne unité alors la cohérence est :

$$\mu(\mathbf{X}) = \max_{i\neq j} |< \mathbf{x}_i, \mathbf{x}_j >|. \tag{1.22}$$

Plus la cohérence est petite moins le design est corrélé. Pour des résultats de convergence sur le Lasso sous des hypothèses de cohérence on pourra se référer à [15], [14].

— la propriété d'isométrie restreinte introduite dans [19]. Elle exige que pour tout sous-ensemble d'indices M de cardinalité bornée par m, le spectre de la matrice ${}^t\mathbf{X}_M\mathbf{X}_M$ (si les colonnes de $\mathbf{X}$ sont de norme euclidienne unité) soit borné par $1 \pm \delta_m$, $\delta_m < 1$. Pour des résultats de convergence du Lasso sous cette hypothèse (techniquement sous une hypothèse plus faible appelée condition de valeur propre restreinte) on pourra se référer à [6].

Sous ce type de conditions, en choisissant un paramètre de régularisation λ de l'ordre de $\sqrt{\log p}$ on retrouve, en supposant α^* k-sparse, une convergence avec une vitesse de l'ordre de $\sigma^2 \log(p)\frac{k}{n}$.

3.3 Les méthodes greedy

Les méthodes greedy sont des heuristiques de résolution itératives du problème de régression parcimonieuse. Elles incorporent les covariables au fur et à mesure, en procédant en une série d'optimisations locales. De nombreuses variantes existent, nous nous contentons donc pour le moment de décrire une forme très générique d'algorithme greedy. On part d'un vecteur de résidus initial $r^0 = \mathbf{y}$, un estimateur initial $\hat{\alpha}^0 = 0$ et un ensemble de covariables initiallement sélectionnées $S^0 = \emptyset$. Supposons qu'on a construit r^{n-1}, $\hat{\alpha}^{n-1}$ et S^{n-1} sans avoir atteint notre critère d'arrêt. Alors :

1. on attribue à chaque covariable du design $\mathbf{X}$ un score basé sur sa proximité avec le vecteur de résidu r^{n-1}.
2. on sélectionne la covariable ayant le plus haut score et on l'ajoute à l'ensemble S^{n-1} pour former S^n.
3. on construit une nouvelle approximation, $\tilde{\mathbf{y}}$, du signal $\mathbf{y}$, dans l'espace vectoriel engendré par les covariables sélectionnées dans S^n.
4. on met à jour les résidus $r^n = \mathbf{y} - \tilde{\mathbf{y}}$ et on itère la procédure.

Pour spécifier une méthode greedy en particulier reste à préciser la façon dont on évalue la proximité entre r^{n-1} et les covariables à l'étape d'attribution du score, et la façon dont on construit l'approximation $\tilde{\mathbf{y}}$. Différents choix mènent aux différentes variantes d'algorithmes greedy. Pour un panorama complet on pourra consulter [88].

On se concentrera ici sur une forme particulière d'algorithme greedy appelée Orthogonal Matching Pursuit (OMP), introduite dans [30] et [79]. Dans cette variante le score de la covariable i au temps $n-1$ est la valeur absolue de son produit scalaire avec r^{n-1}, $| < \mathbf{x}_i, r^{n-1} > |$. Quant au processus d'approximation, il consiste à prendre pour $\tilde{\mathbf{y}}$ le projeté orthogonal de $\mathbf{y}$ sur l'espace vectoriel engendré par les covariables indéxées par S^n. Cette procédure a l'avantage de sa simplicité et de sa rapidité d'exécution.

Tout comme dans le cas de l'estimateur Lasso, l'obtention de bons résultats de convergence pour OMP, sous l'hypothèse de sparsité de α^*, requiert des conditions sur le design $\mathbf{X}$, le même type de conditions que pour le Lasso. Ainsi l'on peut étudier OMP sous des hypothèses de cohérence, comme dans [90] qui fournit une condition suffisante sur le design $\mathbf{X}$ pour avoir une reconstruction parfaite de α^* dans le cas où il n'y pas de terme de bruit, ou encore [52]. Des résultats existent aussi sous des conditions de type RIP comme dans [102]. Tout ces travaux garantissent que la solution obtenue par k pas de OMP est essentiellement comparable à la meilleure approximation à k termes de α^*, [26]. Enfin le critère d'arrêt doit être adapté au terme de bruit dans le modèle 1.2. Une telle étude est conduite dans [16].

4 Etendre la méthode de seuillage aux modèles en grandes dimensions : la méthode one-step greedy

La méthode classique du seuillage 3.1 fonctionne sous l'hypothèse d'orthogonalité du design et n'est donc pas a priori adaptée aux modèles en grandes dimensions. Les méthodes de relaxation convexe 3.2 ou les méthodes greedy 3.3 permettent de retrouver des résultats théoriques comparables aux méthodes de seuillage sous une hypothèse de quasi-orthogonalité du design, ce qui permet de dépasser le cas $p = n$. Mais cette propriété a un coût : l'obtention d'un estimateur par ces deux méthodes peut demander beaucoup plus de calculs qu'un simple seuillage. Il est alors naturel de se demander si les méthodes de seuillage ne peuvent pas être directement adaptées pour s'appliquer aux modèles en grandes dimensions, sous une hypothèse de quasi-orthogonalité du design, c'est-à-dire s'il est possible de pousser la théorie du seuillage au délà de la condition d'orthogonalité, ce qui fournirait une méthodologie au coût computationnel très faible tout en étant efficace.

Cette avancée a été décrite dans une série d'articles ([63], [73], [74], [72]) où une méthode appelée LOL (Learning Out of Leaders) étend le seuillage classique aux designs de cohérence assez faible dans le cas d'un bruit blanc gaussien. Cette théorie est décrite dans la section 4.1. Dans cette thèse, la méthode LOL est adaptée aux bruits gaussiens colorés dans le chapitre 2, et une description de la méthode est développée dans la section 4.2. Enfin la section 4.3 décrit une généralisation des méthodes gloutonnes constituant un cadre général qui contient à la fois les algorithmes greedy au sens de 3.3 et les méthodes comme LOL. La section 4.4 décrit un des articles de cette thèse (restrancrit au chapitre 3) qui discute de l'implémentation pratique d'une telle stratégie.

4.1 Le cas homoscédastique : la méthode LOL (learning out of leaders)

En une série d'articles ([63], [73], [74], [72]) la méthode de seuillage a été étendue aux modèles en grandes dimensions sous des hypothèses de cohérence sur le design avec un bruit blanc. On considère donc ici le modèle 1.2 où $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$ et on portera une attention particulière au cas $p \geq n$ (cette hypothèse n'est pas nécessaire à LOL qui peut aussi être utilisé dans le cas de modèles où le nombre d'observations est supérieur au nombre de covariables, mais LOL est avant tout pensé avec le modèle en grandes dimensions comme principale cible).

On peut résumer la stratégie de seuillage dans le cas d'un design orthogonal à deux étapes : tout d'abord calculer l'estimateur des moindres carrés de α^*, puis le seuiller, c'est-à-dire essentiellement remplacer toutes ses coordonnées plus petites qu'un certain seuil par 0. Cette stratégie, si elle est appliquée telle quelle, est condamnée à l'échec si le design a plus de covariables que d'observations. Nous avons en effet déjà décrit en quoi un estimateur des moindres carrés est inadapté à cette situation.

Pour s'adapter à cette situation on peut raisonner de la sorte : si α^* est sparse alors seul un petit nombre de covariables est important. On peut donc essayer de les sélectionner a priori et travailler sur un modèle réduit, où seules les covariables sélectionnées sont utilisées. On s'est alors ramené à un modèle "classique", où $n \geq p$, et l'on peut procéder au calcul de l'estimateur des moindres carrés et à son seuillage. C'est le principe d'une procédure en deux étapes, qualifiée de sélection / estimation dans [47].

Pour spécifier totalement la méthode LOL il est alors nécessaire de préciser deux points :

1. comment effectuer la sélection intiale des covariables utiles ?
2. à quel niveau doit-on seuiller l'estimateur des moindres carrés final (c'est-à-dire l'estimateur des moindres carrés restreint aux covariables sélectionnées à la première étape) ?

Discutons dans un premier temps de la procédure de sélection des covariables. C'est cette étape qui différentie profondément la méthodologie LOL du seuillage classique, conceptuellement et dans la technique nécessaire aux preuves. L'idée est en fait assez simple, si jamais le design était orthonormal alors l'estimateur $\hat{\alpha} = {}^t\mathbf{X}\mathbf{y}$ (qui n'est autre que l'estimateur des moindres carrés ici) est un estimateur sans biais du vecteur d'intérêt α^*. Si maintenant on ne suppose plus le design orthonormal, mais si on suppose juste que ses colonnes sont normées (ce qui est toujours possible) alors $\hat{\alpha} = {}^t\mathbf{X}\mathbf{y}$ n'est plus un estimateur sans biais de α^*, mais sous de bonnes hypothèses de cohérence il l'est presque ! En effet pour tout $1 \leq i \leq p$ on a que la i-ème coordonnée de $\hat{\alpha}$ vérifie :

$$\hat{\alpha}_i = \alpha_i^* + \underbrace{\sum_{j \neq i} < \mathbf{x}_i, \mathbf{x}_j > \alpha_j^*}_{R_i} + \tilde{\varepsilon}, \tag{1.23}$$

où $\tilde{\varepsilon}$ est une variable gaussienne, $\tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$. On voit alors clairement que dans ce cas $\hat{\alpha}_i$ est un estimateur biaisé de α_i^*, le terme R_i venant biaiser l'estimation. Le terme R_i vient en fait de la corrélation interne du design $\mathbf{X}$, et on peut toujours le majorer à l'aide du concept de cohérence 1.22. En effet on a, pour tout $1 \leq i \leq p$, $|R_i| \leq \mu(\mathbf{X})\|\alpha^*\|_1$. Donc si la cohérence du dictionnaire $\mathbf{X}$ est faible, en particulier si elle est assez faible pour que le terme de biais soit de l'ordre du terme de bruit $\tilde{\varepsilon}$, $\hat{\alpha}$ est un bon estimateur initial de α^* ; il est alors naturel de sélectionner les covariables d'intérêt comme étant celles où $\hat{\alpha}$ est "grand". Cette idée est centrale à la procédure de sélection de LOL, et est en fait utilisée dans de nombreuses procédures de sélections pour les modèles en grandes dimensions. Elle jouit généralement de bonnes propriétés, comme par exemple la notion de Sure Independant Screening de [46]. LOL décide donc de sélectionner toutes les covariables telles que la valeur de $|\hat{\alpha}_i|$ dépasse un certain seuil λ_1.

Il faut alors faire attention à ce que le nombre de covariables sélectionnées ne soit pas trop grand afin de garantir l'unicité de l'estimateur des moindres carrés calculé à l'étape

suivante. La cohérence fournit à nouveau une borne (via le théorème de Greshgorin) sur le nombre de covariables que l'on peut sélectionner tout en garantissant la non-singularité de la matrice de design réduite. En effet on prouve que toute sous-matrice de $\mathbf{X}$ où l'on a conservé au plus $\lfloor \nu/\mu(\mathbf{X}) \rfloor + 1$ colonnes, avec $\nu \in (0,1)$, est non-singulière. LOL sélectionne donc moins de covariables que cette borne (qui est calculable sur des données réelles contrairement à une condition de type RIP [20]).

Une fois les covariables sélectionnées on est ramené alors à un problème de seuillage "classique". Toute la technique consiste alors à exhiber une stratégie universelle de seuillage dans l'étape de sélection qui garantisse (au moins avec grande probabilité) que l'on a choisit les "bonnes" covariables, la détermination du deuxième seuil λ_2 utilisé au moment de régulariser l'estimateur des moindres carrés étant plus classique.

Précisément, notons $B_0(S, M)$ la boule des vecteurs de $\mathbb{R}^p$ dont la sparsité est inférieure à S et dont la norme ℓ_1 est inférieure à M. Alors [63] prouve le théorème suivant.

Théorème 1.5. *Supposons que $p \leq e^{cn}$, pour une certaine constante $c > 0$. Alors si le design $\mathbf{X}$ vérifie que $\mu(\mathbf{X}) \lesssim \sqrt{\frac{\log p}{n}}$, l'estimateur produit par LOL, $\hat{\alpha}^{LOL}$, en choisissant λ_1 et λ_2 de l'ordre de $\sqrt{\frac{\log p}{n}}$ vérifie :*

$$\sup_{\alpha \in B_0(S,M)} \frac{1}{n} \mathbb{E} \|\mathbf{X}\alpha - \mathbf{X}\hat{\alpha}^{LOL}\|_2^2 \lesssim S \frac{\log p}{n}, \tag{1.24}$$

tant que $S < \frac{\nu}{\mu(\mathbf{X})}$, $\nu \in (0,1)$.

Ce qui prouve le caractère minimax de la méthode sur la classe des vecteurs sparses pour des design dont la cohérence est assez faible (sous une hypothèse supplémentaire sur la norme ℓ_1 du vecteur de paramètres). En fait, dans [63] le caractère minimax de la méthode est étendu à toutes les boules ℓ_q, pour $q \in [0,1]$, en passant par un résultat encore plus fort en probabilité (et pas directement en espérance). Ainsi la méthode LOL étend bel et bien le seuillage classique en fournissant toujours une procédure optimale.

4.2 Extension à un bruit coloré

Un des travaux de cette thèse, reproduit au chapitre 2, a consisté à étendre la méthodologie LOL au cas d'un bruit hétéroscédastique, dans l'esprit de l'extension de la méthode de seuillage aux modèles à design orthogonal et à bruit coloré effectuée dans [58], [57]. Une telle extension n'a rien d'immédiat car une technique clé dans la preuve des vitesses de convergence de LOL ne peut plus être utilisée.

En effet lorsqu'on considère un estimateur des moindres carrés restreint à une certaine sous-famille de covariables, le bruit apparait dans l'erreur finale comme le carré de la norme ℓ_2 de la projection orthogonale du terme de bruit initial $\boldsymbol{\varepsilon}$, $\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2$, sur le sous-espace vectoriel V_S de $\mathbb{R}^n$ engendré par les colonnes sélectionnées du design. Or dans le cas

homoscédastique, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I_n)$, et lorsqu'on ne sélectionne pas trop de covariables, $\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2$ est un χ^2 à $|S|$ degrés de liberté dont on peut contrôler la déviation et l'espérance qui ne dépendent donc que du **nombre** de covariables sélectionnées.

Si maintenant on considère, comme dans le chapitre 2, que le bruit est hétéroscédastique, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \Gamma)$, où Γ est une matrice positive définie quelconque, alors la situation change radicalement. En effet la quantité $\mathbb{E}\left[\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2\right]$ ne dépend alors plus seulement du nombre de covariables sélectionnées mais bien de leurs **positions** ! On ne peut donc pas se contenter de modifier le deuxième seuillage pour prendre en compte l'hétéroscédasticité comme on le ferait avec un design orthogonal, mais l'on doit prendre en compte ce phénomène dès l'étape de sélection des covariables.

Il est important de noter que l'on désire éviter deux écueils :

1. on ne veut pas borner $\mathbb{E}\left[\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2\right]$ uniformément sur tous espaces V_S de dimension bornée par une certaine constante. En effet procéder de la sorte reviendrait à considérer le modèle hétéroscédastique comme un modèle homoscédastique avec la pire variance possible. On ne tiendrait pas compte du fait qu'une procédure de sélection efficace tend à sélectionner les covariables du support de α^*, et donc que le terme $\mathbb{E}\left[\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2\right]$ doit être proche de $\|P_{V_{S^*}}[\boldsymbol{\varepsilon}]\|_2^2$, où S^* indexe les coordonnées support de α^*. Or il est tout à fait possible que le signal soit supporté sur une région de l'espace de variance très faible, et par conséquent que $\|P_{V_{S^*}}[\boldsymbol{\varepsilon}]\|_2^2$ soit très inférieur à la pire projection possible.
2. on veut éviter de transformer notre modèle de manière à rendre le bruit blanc. En effet on pourrait multiplier tous les termes de eq. (1.2) par $\Gamma^{-1/2}$ de manière à obtenir un nouveau modèle :
$$\Gamma^{-1/2}\mathbf{y} = \Gamma^{-1/2}\mathbf{X}\alpha^* + \eta,$$
où $\eta \sim \mathcal{N}_n(0, I_n)$. Mais en opérant de la sorte on modifie la matrice de Gram initiale ${}^t\mathbf{X}\mathbf{X}$ en ${}^t\mathbf{X}\Gamma^{-1}\mathbf{X}$ et il y donc un trade-off : tout gain en termes de bruit peut être plus que compensé par une perte de cohérence dans le nouveau design. On veut donc éviter cette opération de "whitening", et développer une méthode utilisant le design original.

De manière à pouvoir contrôler $\mathbb{E}\left[\|P_{V_S}[\boldsymbol{\varepsilon}]\|_2^2\right]$ il est nécessaire d'imposer des restrictions sur le choix de S que notre procédure de sélection impose. Tout d'abord si l'on part de l'estimateur initial $\hat{\alpha} = {}^t\mathbf{X}\mathbf{y}$, il est facile de voir que l'hétéroscédasticité se traduit par :

$$\text{Var } (\hat{\alpha}_l) = \|\Gamma^{1/2}\mathbf{x}_l\|_2^2.$$

Pour des raisons techniques, on définit la quantité $\sigma_l^2 = \text{Var } (\hat{\alpha}_l) \vee 1$ pour tout indice $1 \leq l \leq p$, et par extension $\sigma^2(L) = \sum_{l \in L} \sigma_l^2$ qui représente donc essentiellement la variance portée par les colonnes indéxées par L.

Il est alors nécessaire d'introduire les conditions suivantes sur les éléments de $\mathbf{L}^{\lambda}_{\Sigma_*,N}$, la famille des ensembles d'indices sélectionnables :

1. $\forall L \in \mathbf{L}^{\lambda}_{\Sigma_*,N},\ \forall l \in L,\ |\hat{\alpha}_l/\sigma_l| \geq \lambda$,
2. $\sigma^2(L) \leq \Sigma_*$,
3. $|L| \leq N$.

Ainsi dans le cas hétéroscédastique, on cherche à sélectionner les indices i tels que la quantité $|\hat{\alpha}_l|$ renormalisée par sa variance dépasse un certain seuil λ. De plus, comme dans le cas homoscédastique, on doit contrôler le cardinal de l'ensemble des indices sélectionnés mais on doit ici également contrôler sa variance totale $\sigma^2(L)$ (en effet dans le cas homoscédastique ces deux quantités sont proportionnelles et il est donc équivalent de contrôler l'une ou l'autre). Alors une fois que l'on se restreint aux sous-ensembles aléatoires L qui appartiennent à $\mathbf{L}^{\lambda}_{\Sigma_*,N}$ on peut à nouveau contrôler de manière intéressante la quantité $\mathbb{E}\Big[\|P_{V_L}[\boldsymbol{\varepsilon}]\|_2^2\Big]$. En effet dans le chapitre 2 on prouve le résultat suivant.

Proposition 1.6. *Si $L \in \mathbf{L}^{\lambda}_{\Sigma_*,N}$ est un ensemble aléatoire, et s'il existe une constante $\theta > 0$ telle que $\Sigma_* \leq p^{\theta}$, alors*

$$\mathbb{E}\Big[\|P_{V_L}[\boldsymbol{\varepsilon}]\|_2^2\Big] \lesssim \Big(\sigma^2_{\max}(S^*) + \mu(\Gamma^{1/2}\mathbf{X})\Sigma_*\Big) N \log p, \tag{1.25}$$

dès que $\lambda^2 \geq C\Big[(\mu(\mathbf{X})\|\alpha^*\|_1)^2 \vee \frac{\log p}{n}\Big]$ pour une certaine constante $C > 0$. Ici $S^* = \{l; |\alpha^*_l| > \frac{\lambda}{2}\sigma_l\}$ et $\sigma^2_{\max}(S^*) = \max_{l \in S^*} \sigma_l^2$.

La quantité $\mu(\Gamma^{1/2}\mathbf{X})$, que l'on appelle Γ-coherence, tient compte de l'interaction entre le design et la matrice de covariance du bruit, alors que $\sigma^2_{\max}(S^*)$ reflète bien le comportement espéré, l'espérance de la norme de la projection orthogonale du bruit n'étant pas contrôlée par la pire variance possible mais bien par la pire variance du support du signal d'intérêt !

Une fois en possession de cette proposition technique il devient alors possible, en procédant à la sélection des covariables et en contrôlant à la fois le cardinal et la variance totale, d'obtenir la vitesse de convergence de la méthode sur une large classe de boules anisotropes de $\mathbb{R}^p$. Définissons ces boules anisotropes comme :

— pour $q \in (0,1]$, $\mathcal{B}_{q,\sigma}(M) = \Big\{\alpha \in \mathbb{R}^p; \Big(\sum_{l=1}^p \sigma_l^2 |\alpha_l/\sigma_l|^q\Big)^{1/q} \leq M\Big\}$,

— pour $q = 0$, $\mathcal{B}_{0,\sigma}(S, M) = \Big\{\alpha \in \mathbb{R}^p; \sum_{l=1}^p \sigma_l^2 1\{\alpha_l \neq 0\} \leq S, \|\alpha\|_1 \leq M\Big\}$.

Alors sous des hypothèses de cohérence on prouve dans le chapitre 2 le théorème suivant (les conditions techniques sont explicites dans l'article).

Théorème 1.7. *Supposons que la cohérence du design vérifie $\mu(\mathbf{X}) \lesssim \sqrt{\frac{\log p}{n}}$. Alors en choisissant les seuils λ_1 et λ_2 de l'ordre de $\sqrt{\frac{\log p}{n}}$, et si on note $\hat{\alpha}^*$ l'estimateur fournit par la procédure, on obtient que :*

1. *pour tout* $q \in (0,1]$ *:*

$$\forall \alpha \in \mathcal{B}_{q,\sigma}(M), \quad \mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha}^*)\|_2^2\Big] \lesssim \sigma^2_{\max}(S^*)\Big(\frac{\log p}{n}\Big)^{1-q/2}.$$

où S^* *est défini dans la proposition 1.6.*

2. *Si* $S \leq \nu/\tau_n \vee 1$ *:*

$$\forall \alpha \in \mathcal{B}_{0,\sigma}(S, M), \quad \mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha}^*)\|_2^2\Big] \lesssim \sigma^2_{\max}(S^*)\Big(\frac{S\log p}{n}\Big).$$

Ainsi on obtient presque la vitesse minimax du modèle homoscédastique. En effet les vitesses obtenues dépendent maintenant de la "pire" variance portée par le signal α^*. Nous ne savons pas si ces vitesses sont optimales au sens minimax mais elles sont un premier pas dans la compréhension des méthodes glouttonnes pour le modèle linéaire hétéroscédastique en grandes dimensions. En effet le modèle linéaire hétéroscédastique en grandes dimensions a été bien moins étudié que sa version homoscédastique, et alors que des résultats existent pour étendre les méthodes de relaxation convexe à ce cadre [4], [95], [94], [31], [54], il n'y pas eu à notre connaissance d'effort comparable pour les méthodes greedy.

4.3 Les méthodes super greedy

Si l'on compare la méthodologie greedy décrite à la section 3.3 et les méthodologies à un pas décrites aux sections 4.1, 4.2, on constate que dans les deux cas la première opération effectuée consiste à affecter à chaque covariable un score, via le calcul des quantités $|< \mathbf{x}_i, \mathbf{y} >|$ qui est linéaire en p, le nombre de covariables (ce calcul est par contre linéaire ou non en n, en fonction du fait que l'on puisse ou pas utiliser une structure particulière de la matrice de design $\mathbf{X}$, comme la sparsité de ses colonnes). Mais les méthodes greedy mettent à jour à chaque pas leur vecteur de scores, alors que les méthodologies à un pas n'effectuent ce calcul qu'une fois. Ainsi l'operation d'affectation des scores est au pire de complexité $O(np)$ pour un méthode à un pas, alors que pour une méthode greedy qui effectue k pas, elle exige un calcul de complexité $O(knp)$.

Pour les modèles en "très grandes dimensions", où le nombre de covariables est énorme, c'est ce calcul des scores qui souvent domine la complexité totale des procédures greedy. En effet, la méthode d'estimation a généralement un coût quasi-constant, et proportionnel à n, à chaque pas. Il est donc naturel de se demander s'il n'est pas possible d'effectuer moins de calculs du vecteur des scores, dans l'idée des méthodologies à un pas. A l'inverse une méthode à un pas peut ne pas bien se comporter en pratique parce qu'elle ne met pas assez souvent ses scores à jour, et a donc tendance à intégrer de nombreuses covariables dont l'apport en termes de réduction du biais est faible (car trop fortement corrélées à des variables déjà intégrées).

Les méthodes qui cherchent à obtenir le meilleur des deux mondes sont qualifiées de super-greedy dans [67], [66]. Elles procédent comme les méthodes greedy mais à chaque

pas, au lieu d'intégrer une seule covariable, elles en intégrent un nombre fixe, q. Ainsi à sparsité S égale de l'estimateur, alors qu'une méthode greedy évalue S fois son vecteur de scores, une méthode super greedy ne l'évalue que S/q fois. Cette stratégie peut réaliser des économies de temps de calcul considérables, tout en ayant des performances similaires aux méthodes greedy lorsque le design n'est pas trop corrélé. Elles ont aussi souvent des performances supérieures aux méthodes à un pas en pratique, car moins sensibles aux redondances du design. Reste alors la question d'une stratégie adaptative du choix de q.

Une telle méthode super greedy adaptative a été proposée dans [43] où l'on sélectionne, pour un vecteur de score donné, toute les covariables dont le score est supérieur à un certain seuil (comme dans [63]). Mais contrairement à [63], le seuil est choisi en utilisant le principe de False Discovery Rate introduit dans [1], et la procédure est itérée. En pratique une telle procédure repose donc encore sur la connaissance du niveau de bruit. Or dans les modèles en grandes dimensions, l'estimation de ce niveau de bruit s'avère particulièrement difficile.

La stratégie classique pour choisir le nombre d'itérations optimal de OMP est de procéder par validation croisée (ce qu'une bonne implémentation de OMP permet sans surcoût prohibitif). Mais pour une stratégie super greedy, si l'on doit procéder par validation croisée pour à la fois choisir le nombre total de pas k et la taille des pas intermédiaires i, le nombre de couples (i, k) à tester devient vite grand, ce qui induit un temps de calcul important, alors que nous essayons d'accélérer les méthodes greedy !

Le chapitre 3 décrit une méthodologie super greedy adaptative qui ne repose sur aucune connaissance du bruit a priori, et évite de procéder à une validation croisée pour le choix de la longueur des pas intermédiaires. On introduit cette méthode dans la section suivante.

4.4 Rendre adaptative une procédure super greedy : algorithme super greedy avec pivot

Revenons maintenant à OMP et distinguons sa partie d'estimation de sa partie de sélection :

- la méthode d'estimation de OMP consiste, partant d'un score attribué aux covariables (par la méthode de sélection), à insérer à chaque itération la covariable la mieux notée dans l'ensemble des covariables déjà sélectionnées et à calculer l'estimateur des moindres carrés restreint à ce sous-ensemble de colonnes,
- la méthode de sélection, quant à elle, attribue un score à chaque covariables $\mathbf{x}_i$ en calculant $| < \mathbf{x}_i, r > |$, où r est le vecteur courant des résidus.

Comme on l'a déjà remarqué le coût d'un appel à la procédure d'estimation est au pire de l'ordre de $O(np)$. Quant à la méthode d'estimation, une implémentation raisonnable évite de recalculer à chaque étape un estimateur des moindres carrés, sans tenir compte des calculs effectués à l'étape précédente. Pour ce faire supposons que l'on soit au début de

la $(k+1)$-ième itération de la procédure d'estimation. Notons S_k l'ensemble des covariables déjà sélectionnées à l'étape k, et supposons que l'itération précédente nous fournit la factorisation QR du design $\mathbf{X}$ restreint à S_k, $\mathbf{X}_{S_k} = Q_k R_k$. Alors il est facile de mettre à jour cette factorisation en y incorporant une covariable $\mathbf{x}_j$, $j \notin S_k$, $j \in S_{k+1}$, de telle sorte que $\mathbf{X}_{S_{k+1}} = Q_{k+1} R_{k+1}$: par un procédé de Gram-Schmit, il suffit de calculer k produits scalaires, soit une complexité maximale de l'ordre de $O(nk)$ opérations (une description précise de la méthode et son pseudocode est donnée dans le chapitre 3, pour des références générales au calcul matriciel numérique on pourra se reporter à [7] ou [8]). Comme cette procédure est utilisée pour k très inférieur à p, on voit que le coût d'estimation est largement dominé par le coût de sélection.

L'idée des méthodes super greedy est alors de faire appel à la méthode de sélection le moins souvent possible. En effet si l'on se contente de demander à la méthode d'estimation de ne pas insérer de covariables déjà présentes dans son calcul de l'estimateur, il n'est absolument pas nécessaire de faire appel à la méthode de sélection à chaque itération (ce que fait OMP). Ainsi, pour nous, une variante super greedy de OMP peut être identifiée à sa stratégie vis-à-vis de l'appel à la procédure de sélection.

Un cas particulièrement naturel de stratégie de sélection est alors de procéder à la mise à jour des scores toutes les q itérations, pour une certaine constante q (comme dans [67], [66]). Nous noterons ces variantes OMP_q_N, où N désigne le nombre total d'itérations et q la période d'actualisation du vecteur de scores (i.e. le score est mis à jour toutes les q itérations). Ainsi OMP_1_N désigne N itérations de l'algorithme OMP standard, tandis que OMP_N_N désigne une méthode à un pas. La figure 1.2 compare les erreurs de prédiction relative, $\frac{\|\mathbf{y}-\mathbf{X}\hat{\alpha}\|_2^2}{\|\mathbf{y}\|_2^2}$, en fonction du nombre d'itérations effectuées, pour différentes périodes q de mise à jour.

Sur la figure 1.2, on observe les deux extrêmes que sont OMP (soit OMP_1_80 dans notre notation), qui recalcule son vecteur de scores à chaque pas, c'est-à-dire ici 80 fois, et OMP à un pas (OMP_80_80) qui ne le calcule qu'une fois. On observe aussi des stratégies intermédiaires qui nous permettent de constater que :

- une stratégie de périodicité faible, comme OMP_2_80 ou OMP_4_80 ici, se comporte essentiellement comme OMP (car le design est ici très décorrélé), mais ne réalise qu'un gain calculatoire modéré,
- la stratégie à un pas, OMP_80_80, qui réalise un énorme gain de temps de calcul, commet globalement une erreur supérieure à celle de OMP, et a de plus tendance à rencontrer des plateaux, c'est-à-dire des zones où l'adjonction d'une nouvelle covariable ne diminue pas significativement l'erreur de prédiction,
- si l'on observe la stratégie à deux pas, OMP_40_80, on constate l'effet associé à la mise à jour des scores (réalisée au pas 40, et indiquée par un trait vertical sur la figure 1.2) : l'erreur commise se démarque alors de celle de la méthode à un pas, sort du plateau d'erreur, et diminue à nouveau pour se rapprocher de celle de OMP.

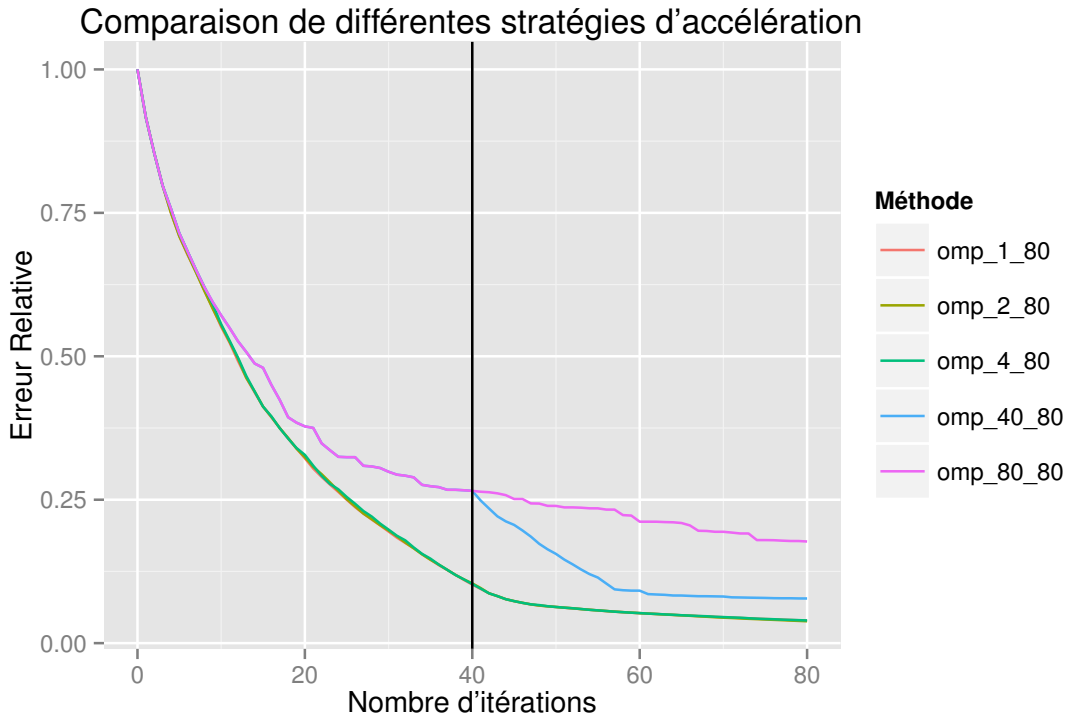

FIGURE 1.2. Comparaison de OMP et de ses formes super greedy. Le design est une matrice à entrées gaussiennes i.i.d, avec $n = 500$ et $p = 1500$. Le paramètre α^* est sparse, de sparsité $S = 50$.

La méthode que nous proposons au chapitre 3 implémente une stratégie adaptative qui essaye de maintenir une erreur de l'ordre de celle commise par OMP, tout en actualisant le moins possible le vecteur des scores. Pour ce faire on utilise le vecteur des résidus actualisé r_k, retourné par la k-ième itération de la procédure d'estimation. On peut donc à chaque itération calculer le ratio $\frac{\|r_k\|_2}{\|r_{k-1}\|_2}$ qui mesure, en proportion, le gain en pouvoir prédictif réalisé par l'adjonction de la variable introduite à l'étape k. On procède alors de la sorte :

1. on part d'un vecteur de score initial,
2. on incorpore les covariables une par une dans l'ordre induit par le vecteur de scores initial, tant que le ratio $\frac{\|r_k\|_2}{\|r_{k-1}\|_2}$ est inférieur à une constante $\lambda \in (0, 1)$,
3. si à une certaine itération k_0 l'adjonction d'une nouvelle covariable ne respecte pas la relation $\frac{\|r_{k_0}\|_2}{\|r_{k_0-1}\|_2} < \lambda$, alors seulement on actualise le vecteur de score et on redémarre la procédure à l'étape $k_0 - 1$.

Ainsi cette procédure ne tend à actualiser son vecteur de scores (on dira qu'elle pivote) que lorsque son erreur de prédiction (estimée sur le training set) rencontre un plateau (i.e. que la norme du vecteur de résidu ne décroît pas suffisament vite). Cette stratégie assure à notre forme super greedy de OMP, d'avoir une erreur de prédiction qui décroît à une vitesse comparable à celle de OMP tout en effectuant peu d'actualisations des scores. On peut se référer à la figure 1.3, pour voir notre méthode en action sur un modèle similaire à celui employé pour l'expérience de la figure 1.2. Les traits verticaux précisent à quelles itérations notre stratégie décide de recalculer les scores.

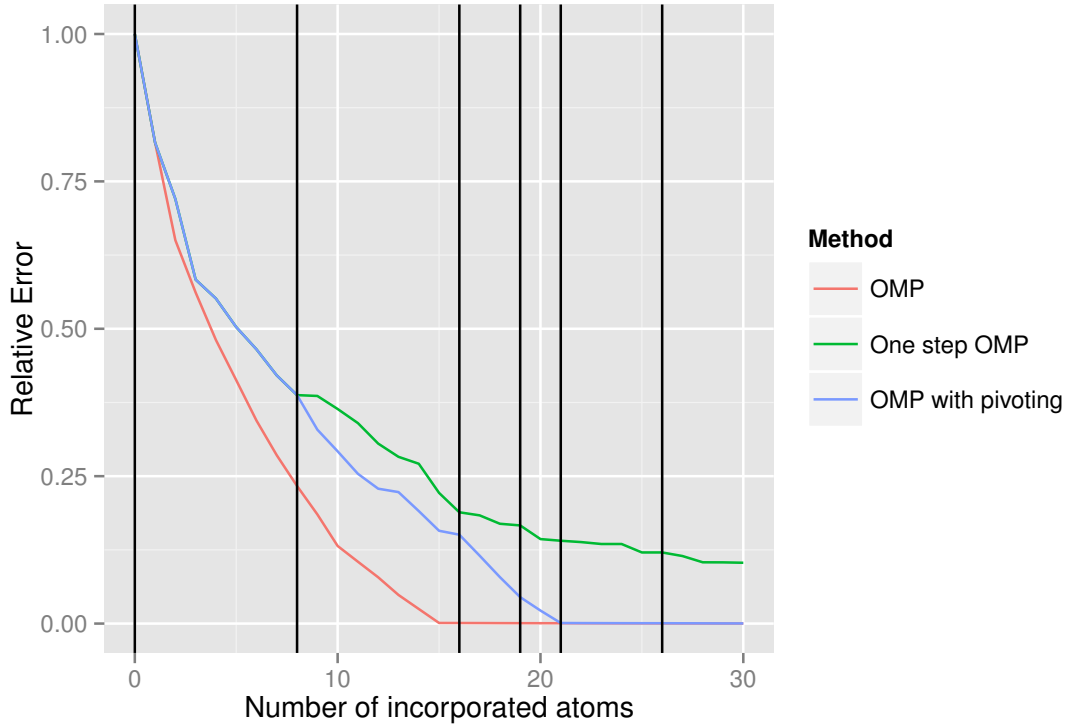

FIGURE 1.3. Comparaison de OMP, OMP à un pas et notre méthodologie avec pivot. Le design est une matrice à entrées gaussiennes i.i.d, avec $n = 75$ et $p = 300$. Le paramètre α^* est sparse, de sparsité $S = 15$. Les lignes verticales indiquent les étapes où la méthodologie avec pivot met à jour le vecteur de score.

Pour illustrer les gains très importants en temps de calcul réalisés par notre méthode, on a mesuré et reporté les resultats sur la figure 1.4, le temps d'obtention d'un estimateur à sparsité fixée (i.e. on fixe le nombre d'itérations) par OMP et par OMP avec pivot, en fonction du nombre de covariables présentes dans le design. Le nombre d'itérations total étant fixé, seul le coût associé à la procédure de sélection différentie les deux méthodes, et on peut constater que cela induit un gain considérable.

De plus le paramètre λ régularise la procédure : en effet plus λ est proche de 0, plus l'agorithme s'arrête rapidement (incapable de réduire la norme du vecteur de résidus dans les proportions demandées), et actualise régulièrement ses scores. A l'inverse, lorsque $\lambda = 1$, alors la procédure devient une méthode à un pas qui se contente du vecteur de scores initial. Il suffit donc de le sélectionner par validation croisée (ici une seule validation croisée suffit au lieu des deux nécessaires a priori dans une méthode de type OMP_q_N).

Enfin un autre avantage de la procédure avec pivot est illustré sur des données réelles au chapitre 3. On y considère les données fournies par la compétition Kaggle `https://www.kaggle.com/c/job-salary-prediction`. On y trouve différentes offres d'emploi, avec leur description et le salaire associé. Au corpus des descriptions on peut associer une matrice "Document-terms" de grandes dimensions, où les documents sont

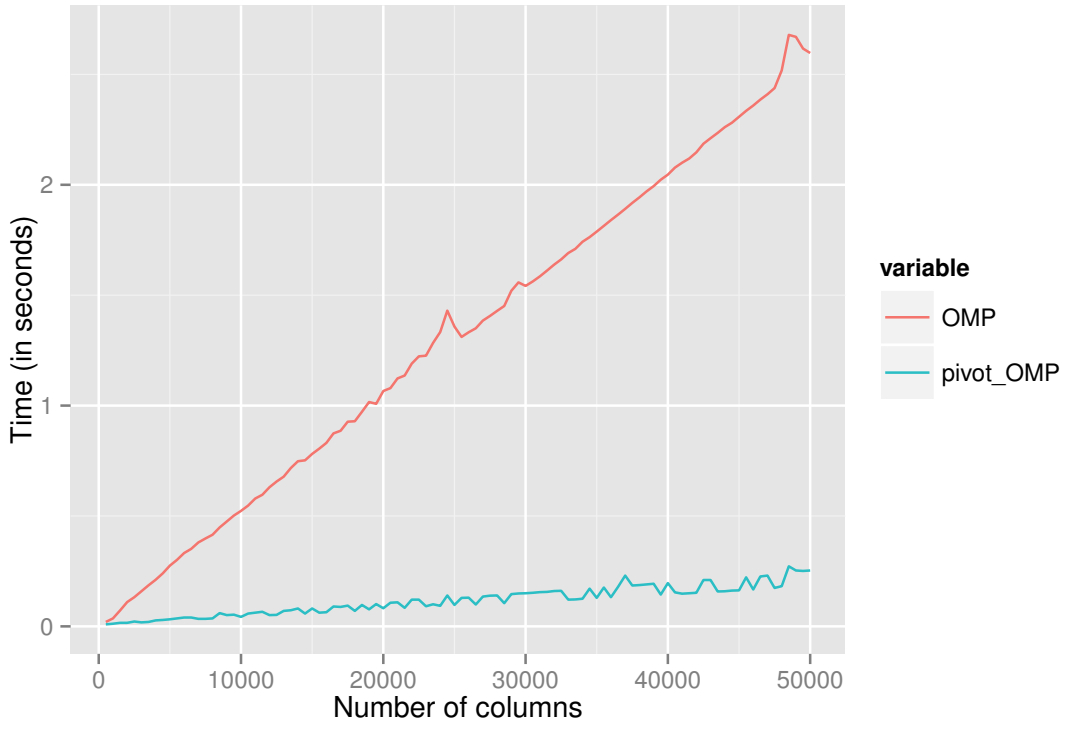

FIGURE 1.4. Comparaison du temps d'éxécution de OMP et de notre méthodologie. Le design est une matrice gaussienne avec un nombre fixe d'observations, $n = 750$. Le paramètre α^* est sparse, de sparsité $S = 50$, les deux méthodes réalisant 150 pas.

en ligne et les termes en colonne, et chaque coefficient vaut 1 ou 0, en fonction du fait qu'un certain terme soit présent, ou pas, dans le document concerné. On cherche alors un estimateur permettant de prédire le salaire à partir de la description de l'emploi.

Pour les comparer, on calcule l'estimateur fournit par OMP et par notre modification avec pivot sur un jeu de données "train" et on mesure leur erreur sur un jeu de données "test" indépendant. On reporte les résultats sur la figure 1.5.

Il est intéressant de constater ici que notre méthode donne toujours de meilleurs résultats que OMP. Il semble raisonnable de supposer que celà vient de la capacité de notre méthode à incorporer une nouvelle covariable même si elle est corréllée à des covariables déjà sélectionnées (du moins dans une certain proportion fixée par λ). Ainsi si l'on compare les termes sélectionnés par les deux méthodes on peut constater bien plus de redondance dans ceux choisis par notre algorithme comparativement à OMP (par exemple on peut voir sur le tableau 1.1 que la méthode avec pivot sélectionne "projects" et "project". Cette corrélation que l'on autorise parmi les covariables sélectionnées permet probablement d'éviter d'incorporer les artefacts que OMP tend à utiliser, en actualisant trop régulièrement les scores.

FIGURE 1.5. Comparaison de l'erreur de prédiction relative, estimée sur un jeu de données indépendant du train, de OMP et de notre variation avec pivot, en fonction du nombre total d'itérations.

	OMP	OMP with pivoting
1	and	and
2	chase	the
3	projects	for
4	ooh	chase
5	business	locum
6	own	projects
7	london	project
8	management	analysis
9	analysis	business
10	paye	technical

TABLE 1.1. Dix premiers termes choisis par OMP et par notre méthode.

5 Estimation non-paramétrique : le cas de l'estimation de densité

L'estimation non-paramétrique diffère de l'estimation paramétrique, dont le modèle linéaire est un bon exemple, en supposant, non pas que le vecteur α^* à estimer appartient à un espace euclidien (même de grande dimension), mais plutôt à un espace de fonctions, c'est-à-dire un espace de dimension infinie. Pourtant, pour l'estimer, on ne dispose toujours que d'un nombre fini d'observations, n. On s'intéresse dans la suite à un modèle particulier

d'estimation non-paramétrique, l'estimation de densité. Pour une introduction générale au sujet (dont s'inspire largement cette section) on pourra consulter [92].

5.1 Estimation de densité sur $\mathbb{R}$ et risque minimax

Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. de densité de probabilité f_X par rapport à la mesure de Lebesgue sur $\mathbb{R}$. Le problème de l'estimation de densité est alors, partant des observations X_i, de construire (de façon mesurable) un estimateur $\hat{f}_n$ de f_X. Ce problème est dit non-paramétrique lorsque l'objet à estimer, f_X, vit a priori dans un espace de dimension infinie, i.e. lorqu'on ne veut pas imposer a priori à f_X d'appartenir à une certaine famille paramétrée de densités !

Supposons que f_X appartienne à une certaine classe non-paramétrique de densités $\mathcal{F}$. Alors, si d est une semi-distance sur $\mathcal{F}$, on peut, comme pour le modèle linéaire, introduire la notion de risque de l'estimateur $\hat{f}_n$:

$$R(\hat{f}_n, f_X) = \mathbb{E}\Big[d^2(\hat{f}_n, f_X)\Big].$$

On peut alors, comme on l'a fait pour le modèle linéaire, qualifier un estimateur $\hat{f}_n$ de minimax sur $\mathcal{F}$ s'il existe une suite positive $(\psi_n)_{n\geq 1}$ telle que le risque maximal sur $\mathcal{F}$ vérifie :

$$r(\hat{f}_n) = \sup_{f_X \in \mathcal{F}} \mathbb{E}\Big[d^2(\hat{f}_n, f_X)\Big] \leq C\psi_n^2,$$

pour une certaine constante $C > 0$, et si le risque minimax sur $\mathcal{F}$:

$$R_n^* = \inf_{\hat{g}_n} \sup_{f_X \in \mathcal{F}} \mathbb{E}\Big[d^2(\hat{f}_n, f_X)\Big],$$

où l'infinimum est pris sur tous les estimateurs $\hat{g}_n$, vérifie

$$\liminf_{n\to+\infty} \psi_n^{-2} R_n^* \geq c,$$

pour une certaine constante $c > 0$.

5.2 Estimateur à noyau d'une densité

Introduit dans [82], puis généralisé dans [78], l'estimateur à noyau est une des méthodes les plus classiques d'estimation d'une densité.

On qualifie de noyau toute fonction $K : \mathbb{R} \to \mathbb{R}$, intégrable, telle que $\int K(u)\mathrm{d}u = 1$. Soit $h > 0$, l'estimateur à noyau, de noyau K et de fenêtre h, de f_X est la fonction,

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big), \tag{1.26}$$

définie pour tout $x \in \mathbb{R}$. Notons, pour tout $h > 0$, $K_h(x) = \frac{1}{h}K(\frac{x}{h})$. Alors, par construction, l'espérance de l'estimateur à noyau eq. (1.26) n'est autre que le produit de convolution de f_X avec K_h :

$$\mathbb{E}[\hat{f}_h(x_0)] = K_h * f_X(x_0) = \int_{\mathbb{R}} K_h(x_0 - y) f_X(y)\mathrm{d}y.$$

L'étude de cet estimateur est le sujet d'une importante littérature, et l'on pourra se référer à [86], [91], [35], [34] pour une introduction à la théorie sous différents points de vue. En particulier l'estimateur à noyau est minimax sur de nombreuses classes de densités (densités appartenant à des classes de Hölder ou à des espaces de Sobolev par exemple). De plus un choix de la fenêtre dépendant des observations (on peut citer en particulier la méthode fondamentale dite de Lepski, [65]) en fait un estimateur adaptatif. Malgré ses nombreuses bonnes propriétés, l'estimateur à noyau souffre de défauts. On s'intéressera en particulier au phénomène dit de "boundary bias", ou biais au bord, décrit dans la section suivante.

5.3 Le phénomène de biais au bord (ou boundary bias)

Supposons que la densité d'intérêt soit supportée par un intervalle I de $\mathbb{R}$ admettant une frontière non vide. Les cas de variables aléatoires modélisant une proportion, et donc à valeurs dans $[0,1]$, ou encore modélisant des temps d'arrivée ou de survie, et donc à valeurs dans $\mathbb{R}^+$, sont particulièrement importants en pratique.

Supposons que la densité f_X appartienne à la classe de Hölder $\Sigma(\beta, L, I)$, avec $\beta, L > 0$, c'est-à-dire à la classe des densités $l = \lfloor \beta \rfloor$ fois dérivables sur I, dont toutes les dérivées sont bornées, et telles que :

$$|f_X^{(l)}(y) - f_X^{(l)}(x)| \leq L|y-x|^{\beta-l}, \ \forall x, y \in I. \tag{1.27}$$

Alors si f_X admet une limite non nulle, à droite ou à gauche, en un point à la frontière de son support, f_X n'est pas globalement lisse sur $\mathbb{R}$, $f_X \notin \Sigma(\beta, L', \mathbb{R})$. Or (et il s'agit d'un résultat classique sur la convolution) si f_X admet une discontinuité en un point x_0, et admet des limites finies de chaque côté de x_0, et si de plus le noyau K est pair (ce qui est presque toujours le cas en pratique), alors :

$$\mathbb{E}[\hat{f}_h(x_0)] = K_h * f_X(x_0) \to \frac{f_X(x_0^+) + f_X(x_0^-)}{2}, \ \text{ lorsque } h \to 0. \tag{1.28}$$

Ainsi l'estimateur à noyau n'est même plus consistant sur le frontière de I !

De plus le biais de l'estimateur à noyau est bien plus grand près du bord que sur les points intérieurs, ce qui induit une détérioration de la vitesse de convergence de l'erreur quadratique moyenne vers 0 sur tous les points trop proches de la frontière. Cette situation est illustrée sur la fig. 1.6.

Plus précisément, fixons par simplicité $I =]-1, 1[$, et supposons que $f_X \in \Sigma(\beta, L,]-1, 1[)$. Alors le biais en $x_0 \in]-1, 1[$, $b(x_0) = \mathbb{E}[\hat{f}_h(x_0)] - f_X(x_0)$, vérifie, dès que le noyau

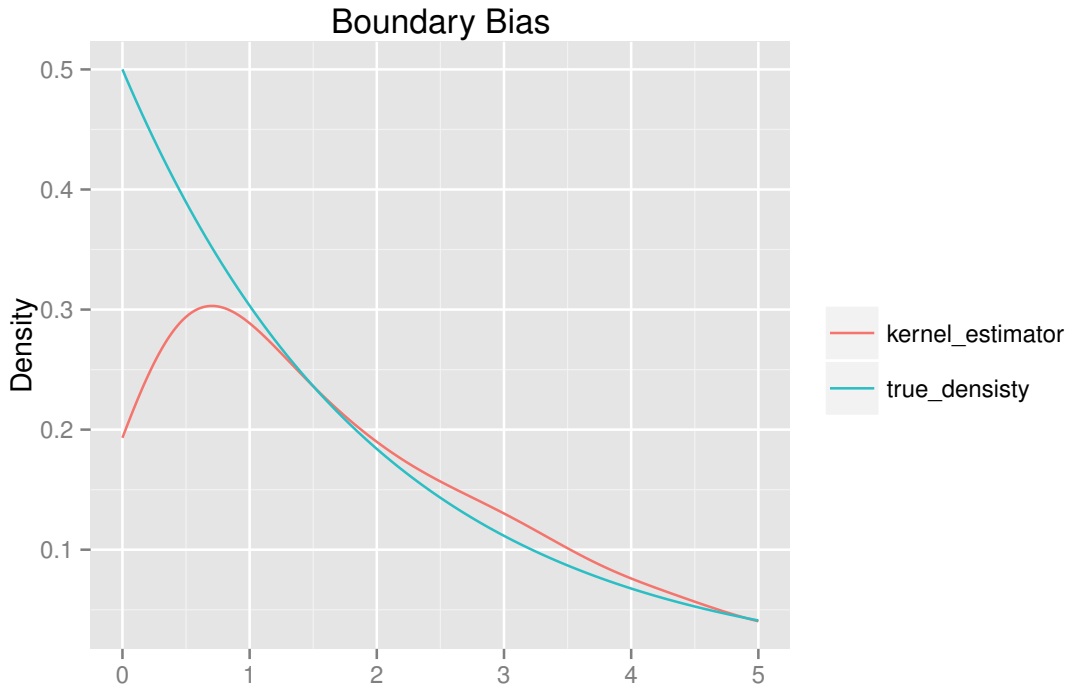

FIGURE 1.6. Estimation d'une densité exponentielle de paramètre 0.5 avec un noyau gaussien, à partir de $n = 1000$ observations.

K est supporté sur $[-1, 1]$:

$$b(x_0) = f(x_0)(t_0(x_0, h) - 1) + \sum_{k=1}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0) t_k(x_0, h) h^k$$
$$+ \frac{(-1)^l}{l!} h^l \int_{\frac{x_0-1}{h} \vee -1}^{\frac{x_0+1}{h} \wedge 1} t^l K(t) f^{(l)}(x_0 - \tau t h) \mathrm{d}t,$$

où $t_k(x, h) = \int_{\frac{x-1}{h} \vee -1}^{\frac{x+1}{h} \wedge 1} t^k K(t) \mathrm{d}t$, pour tout $x \in]-1, 1[$, et $h > 0$.

On distingue alors :

— les points intérieurs, x_0, tels que $-1 + h \leq x_0 \leq 1 - h$ (on supposera que $h < 1$, afin de garantir leur existence). En ces points on a :

$$t_k(x_0, h) = \int_{-1}^{1} t^k K(t) \mathrm{d}t,$$

et l'on sait construire des noyaux K (voir par exemple la construction de [91] utilisant des polynômes de Legendre) d'ordre l, pour tout entier l, tels que $t_0(x_0, h) = 1$ et $t_k(x_0, h) = 0$ pour tout $1 \leq k \leq l$. On en déduit que le biais se comporte, pour une densité $f_X \in \Sigma(\beta, L,]-1, 1[)$, comme $O(h^\beta)$ en tout point intérieur (ce qui permet d'atteindre la vitesse minimax).

— les points au bord, i.e. tels que $-1 \leq x_0 < -1 + h$ ou $1 - h < x_0 \leq 1$, pour lesquels il est nécessaire de modifier notre estimateur si l'on veut à nouveau obtenir un biais en $O(h^\beta)$.

Ce phénomène a d'importantes conséquences dans de nombreuses situations : en effet il induit le praticien à sous-estimer la probabilité pour la variable X, de densité f_X supportée sur un intervalle I, à prendre des valeurs proches de la frontière de I. Corriger cette sous-évaluation a été le sujet d'une importante activité de recherche dont nous décrivons quelques grands axes. On considérera à partir de maintenant les points du bord droit par simplicité, i.e. les points x_0 tels que $1-h < x_0 \leq 1$. Il est également plus aisé de les reparamétrer en les notant $x_\alpha = 1 - \alpha h$, pour $\alpha \in (0,1)$, de telle sorte que $t_k(x_\alpha, h) = \int_{-\alpha}^{1} t^k K(t)\mathrm{d}t$.

Estimation consistante au bord : la méthode de réflexion et la méthode "cut-and-normalize". Les premières méthodes de correction du phénomène de boundary bias pour l'estimateur à noyau avaient pour but de le rendre consistant, c'est-à-dire de garantir que $t_0(x_0, h) = 1$ en tout point x_0 du bord.

La méthode de réflexion [84], [86], [25] consiste à ajouter à nos observations leurs réflexions par rapport aux points de la frontière, et à construire à partir de ces observations augmentées un estimateur à noyau $\hat{f}_n^*$. On peut alors en déduire un estimateur de f_X en considérant $\hat{f}_n = C\hat{f}_n^* \mathbb{1}_I$, où C est une constante de renormalisation.

La méthode dite de "cut and normalize", [50], considère au point x_α, le noyau :

$$K_\alpha(t) = \frac{1}{\int_{-\alpha}^{1} K(t)\mathrm{d}t} K(t)\mathbb{1}\{-\alpha \leq t \leq 1\}.$$

On peut prouver que ces deux méthodes sont en fait équivalentes et fournissent un estimateur consistant de f_X en -1 et 1. Par contre, elles ne garantissent qu'un biais en $O(h)$, même lorsque f_X est plusieurs fois dérivables (sauf à adjoindre l'hypothèse $f_X'(1) = f_X'(-1) = 0$, qualifiée de "shoulder condition", pour obtenir un biais en $O(h^2)$).

La méthode de transformation. La méthode de transformation introduite dans [35], puis étudiée dans [96], [83] propose de transformer les observations à l'aide d'un $\mathcal{C}^1$ difféomorphisme $\Phi^{-1} : I \to \mathbb{R}$ afin de se ramener à une densité sur $\mathbb{R}$, d'estimer cette densité puis de la retansformer pour obtenir un estimateur final de f_X. Ainsi l'estimateur "back-transformed" s'écrit :

$$\hat{f}_n^T(x_0) = \frac{1}{nh\phi(\Phi^{-1}(x_0))} \sum_{i=1}^{n} K\Big(\frac{\Phi^{-1}(x_0) - \Phi^{-1}(X_i)}{h}\Big),$$

où ϕ est la dérivée de Φ. On dispose en fait d'assez peu de résultats théoriques sur cet estimateur, la méthodologie de [96] reposant sur une famille paramétrée de transformations. Pour obtenir des résultats plus fins, avec un biais de l'ordre de $O(h^2)$ pour des densités deux fois dérivables, [51] procède plus subtilement en utilisant une méthode de vraisemblance locale pour estimer la densité transformée.

Noyaux Beta et Gamma. Dans une série d'articles [23], [12], [22], [21] il a été proposé d'utiliser des noyaux basés sur la loi Beta (pour des densités à support compact) ou Gamma (pour des densités à support positif). Par exemple si f_X est supportée sur $[0,1]$, l'estimateur à noyau Beta est :

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x/b+1,(1-x)/b+1}(X_i),$$

où $K_{p,q}$ est la densité de la loi Beta(p,q), et où le paramètre b joue le rôle de la fenêtre. Mais là encore pour obtenir un biais en $O(h^2)$ dans le cas de l'estimateur à noyau Gamma, [99] prouve que la shoulder condition est nécessaire, alors que [5] prouve que l'estimateur à noyau Beta n'est minimax que lorsque la régularité de f_X est inférieure à 2.

La méthode des boundary kernels. La méthode des boundary kernels, introduite dans [50], cherche à associer à chaque point x_α un noyau K_α de biais optimal. On pourra consulter [75], [61], [60], [100], [101]. Ainsi on qualifiera de boundary kernel d'ordre $2l+1$, $l \geq 0$, au point x_α, $\alpha \in (0,1)$, tout noyau K_α tel que :

$$t_k(x_\alpha, h) = \int_{-\alpha}^{1} t^k K_\alpha(t)\mathrm{d}t = \delta_{0k},\ \forall\, 0 \leq k \leq 2l+1. \tag{1.29}$$

La technique des boundary kernels a l'avantage immédiat de ne pas se limiter à l'ordre 2 mais est adaptable à tout ordre. Dans [49], l'existence de boundary kernels de tout ordre et de norme L^2 minimale est demontrée (en effet la variance d'un estimateur à noyau en un point est proportionnelle à la norme L^2 du noyau). Ils apparaissent comme solutions d'un système linéaire $(2l+1)\times(2l+1)$ ([75] généralise la construction).

On s'est ici uniquement concentré sur quelques approches parmi les plus étudiées mais de nombreuses autres existent (par exemple la méthode de pseudo-données [27], ou encore les méthodes par diffusion [9]). De plus certaines méthodes d'estimation de la densité ne manifestent pas le phénomène de boundary bias, comme les méthodes de polynômes locaux [24], [45].

Nous présentons au chapitre 4 une construction nouvelle d'un boundary kernel, qui se contente de modifier, en chaque point x_α, un noyau d'ordre $2l+1$ adapté aux points intérieurs pour le transformer en boundary kernel d'ordre $2l+1$ en x_α. Le fait de partir d'un noyau intérieur et de le modifier semble être plus légitime pour l'utilisateur, habitué à choisir son noyau de prédilection pour les points intérieurs, à qui on n'impose pas l'utilisation d'un noyau complétement différent au bord (ce que ferait une méthode du type de celle de [49], en imposant de plus des conditions assez arbitraires sur le noyau). De plus la construction est simple, rapide (elle repose sur la résolution d'un système $(l+1)\times(l+1)$), et s'adapte facilement à tout ordre du noyau initial.

5.4 Modification de noyaux d'ordre quelconque au bord

Si l'on reprend l'argument qui mène à la convergence 1.28 (on pourra par exemple consulter [62]), on voit que la parité du noyau K y joue un rôle central. Considérons l'effet d'une convolution par un noyau impair $\tilde{K}$. Pour éviter des cas triviaux (comme $\tilde{K} = 0$) on supposera de plus que notre noyau vérifie $\int_{\mathbb{R}^+} \tilde{K}(u)\mathrm{d}u = P(\tilde{K}) \neq 0$. Alors il est facile de vérifier (sous des hypothèses techniques similaires au cas pair) que :

$$\tilde{K}_h * f_X(x_0) \to P(\tilde{K})\Big(f_X(x_0^-) - f_X(x_0^+)\Big), \text{ lorsque } h \to 0. \tag{1.30}$$

Ainsi, partant d'un noyau pair K, si on lui ajoute une fonction impaire $\tilde{K}$ (ce qui ne modifie pas sa "masse" totale), telle que $P(\tilde{K}) = 1/2$ (pour le bord droit, pour le bord gauche il faut $P(\tilde{K}) = -1/2$), pour former un nouveau noyau $\bar{K} = K + \tilde{K}$, on récupère un estimateur à noyau consistant au bord.

Considérons un noyau initial K d'ordre $2l+1$ à l'intérieur, supporté sur $[-1, 1]$. On va chercher à construire, pour tout $\alpha \in (0, 1)$, une fonction impaire $\tilde{K}_\alpha$, elle aussi supportée sur $[-1, 1]$, telle que $K_\alpha(u) = K(u) + C_\alpha \tilde{K}_\alpha(u)$ est un boundary kernel (à droite) du même ordre que K, c'est-à-dire vérifie :

$$t_k(x_\alpha, h) = \int_{-\alpha}^{1} t^k K_\alpha(t)\mathrm{d}t = \delta_{0k}, \ \forall\, 0 \leq k \leq 2l+1, \tag{1.31}$$

et où la constante C_α est telle que $C_\alpha \int_0^1 \tilde{K}_\alpha(u)\mathrm{d}u = 1/2$. Cette approche se distingue de l'approche dite de "generalized jackkniffe", [61], qui fixe un second noyau $\tilde{K}$ et cherche un à corriger le biais au bord en x_α en considérant un noyau de la forme $\bar{K} = a_\alpha K + b_\alpha \tilde{K}$, où les constantes a_α, b_α sont des fonctions du points x_α. Ici on modifie directement le noyau $\tilde{K}_\alpha$ pour s'affranchir du biais au bord, ce qui a pour avantage de facilement se généraliser à n'importe quel ordre de correction (et pas seulement à l'ordre 2).

Pour ce faire, étant donné $\alpha \in (0, 1)$, on procède en deux étapes. Tout d'abord on va "replier" le noyau K sur $[\alpha, 1]$. Ainsi on définit d'abord $\tilde{K}_\alpha$ pour tout $u \in [\alpha, 1]$ en symétrisant le noyau K :

$$\tilde{K}_\alpha(u) = K(u), \ \forall u \in [\alpha, 1]. \tag{1.32}$$

Cette relation garantit à K_α d'avoir tout ses moments t_k pairs vérifiant bien la relation 1.31. La figure 1.7 illustre cette opération de pliage (en choisissant pour le moment de fixer $\tilde{K}_\alpha(u) = 0$, pour tout $u \in [0, \alpha)$). Cette opération de symmétrisation est différente de la technique de réflexion de [84], en effet elle ne modifie pas le noyau K sur $[0, \alpha)$, et pourtant elle est plus puissante puisqu'elle permet d'obtenir la nullité de tous les moments pairs au bord et pas seulement du moment d'ordre 0.

Reste alors à définir $\tilde{K}_\alpha$ pour tout $u \in]\alpha, 1]$, de manière à assurer la nullité des moments au bord impairs en x_α de K_α. Pour ce faire il est naturel de prolonger la technique de [49],

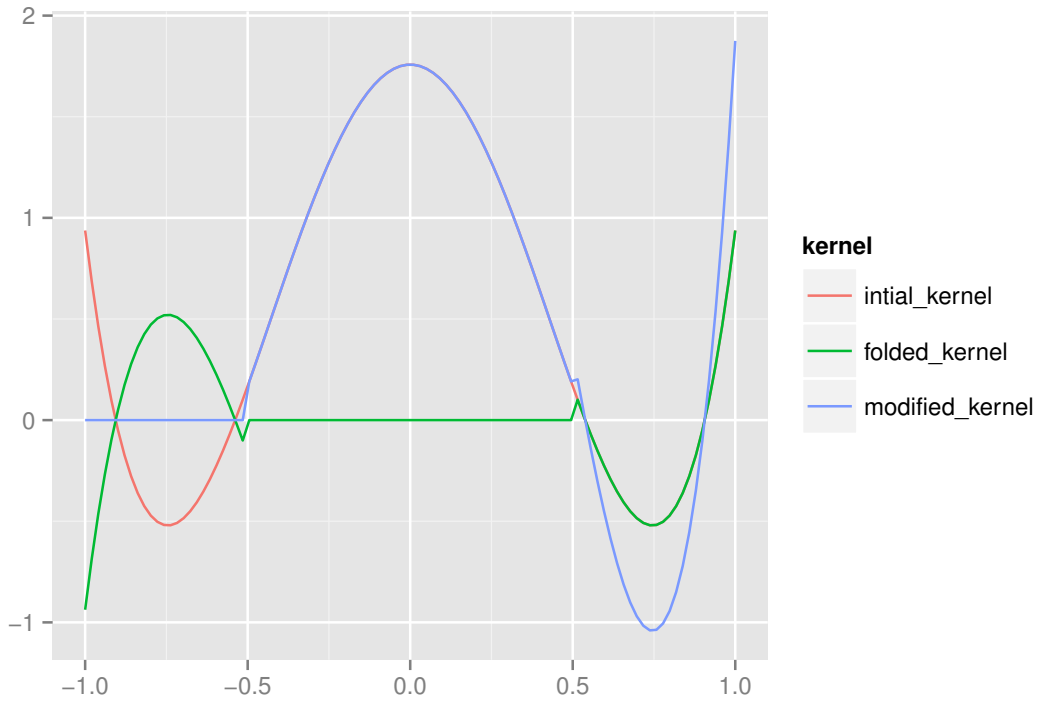

FIGURE 1.7. Illustration de l'opération de "pliage" d'un noyau initial, ici un noyau d'ordre 5 (en rouge), au point où $\alpha = 1/2$.

en cherchant $\tilde{K}_\alpha$ sous la forme d'un polynôme sur cet intervalle. Par contre ici la condition de symétrie sur $\tilde{K}_\alpha$ nous permet de nous contenter de chercher un polynôme de degré l :

$$\tilde{K}_\alpha(u) = \sum_{j=0}^{l} a_j u^j, \ \forall u \in [0, \alpha[.$$

Alors en injectant cette expression dans le calcul des moments de bord, et en leur imposant la relation 1.29, on obtient que les scalaires a_j doivent être solutions du système linéaire :

$$\Lambda \mathbf{a} = \mathbf{m}(\alpha), \tag{1.33}$$

où Λ est une matrice $(l+1) \times (l+1)$ de terme général $\Lambda_{kj} = \frac{\alpha^{2k+j+2}}{2k+j+1}$, et où $m_k(\alpha) = \int_{-\alpha}^{1} u^{2k+1} K(u) \mathrm{d}u$. En effet de simples considérations de symétrie imposent à la fonction $\tilde{K}_\alpha$ de vérifier pour tout $0 \leq k \leq l$:

$$t_{2k+1}(x_\alpha, h) = 2m_k(\alpha) + 2\int_0^\alpha u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u, \tag{1.34}$$

ce qui implique la relation 1.33. On démontre (on peut se référer au chapitre 3) que ce système est toujours inversible pour $\alpha \in (0, 1]$ (bien entendu lorsque $\alpha = 0$ la correction $\tilde{K}_0$ n'est autre que le noyau K replié).

Ainsi le calcul de la correction au bord pour un noyau d'ordre $2l+1$ se réduit, en utilisant la relation de "pliage" 1.32, à la résolution d'un système $(l+1) \times (l+1)$.

On peut calculer explicitement ce boundary kernel dans de nombreux cas. Par exemple considérons le noyau d'ordre 3, $K(u) = \left(\frac{9}{8} - \frac{15}{8}u^2\right)\mathbb{1}\{|u| \leq 1\}$, alors on vérifie que $m_0(1/2) = -9/512$ et $m_1(1/2) = -135/3072$. Il est alors nécessaire de résoudre le système :

$$\begin{pmatrix} 1/4 & 1/16 \\ 1/48 & 1/128 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} -9/512 \\ -135/3072 \end{pmatrix} \tag{1.35}$$

Ce qui donne pour $u \in [0, 1/2[$, $\tilde{K}_\alpha(u) = 4.007 - 16.312u$.

Il est important de noter, pour la construction pratique de ces noyaux, que bien que le système 1.33 soit inversible, il est très mal conditionné numériquement.

Pour pallier à cette difficulté considérons la famille des polynômes de Legendre ([87], [3]) $(L_k)_{k\geq 0}$, qui forme une famille orthonormale de $L^2([-1, 1])$ (muni du produit scalaire usuel). Alors pour tout $k \geq 0$, les applications :

$$\tilde{L}_k : x \mapsto \frac{1}{\sqrt{\alpha}} L_k\Big(\frac{x}{\alpha}\Big),\ \forall\, x \in [-\alpha, \alpha],$$

forment une famille orthonormale de $L^2([-\alpha, \alpha])$. Alors on peut chercher la restriction de $\tilde{K}_\alpha$ à $[-\alpha, \alpha]$ sous la forme :

$$\tilde{K}_\alpha(u) = \sum_{i=0}^{l} \beta_i \tilde{L}_{2i+1}(u),\ \forall\, u \in [-\alpha, \alpha]. \tag{1.36}$$

Comme précédemment, cette écriture induit un système linéaire 1.33, où la matrice $\Lambda(\alpha)$ a pour terme général $\Lambda_{ij}(\alpha) = \alpha^{2i+1/2} \int_0^1 L_{2j+1}(y) y^{2i+1} \mathrm{d}y$.

Ce système est bien mieux conditionné que le précédent et permet le calcul explicite des fonctions K_α et $\tilde{K}_\alpha$. Ainsi partant du noyau d'Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \leq 1\}$, la figure 1.8 représente la fonction de correction $\tilde{K}_{1/2}$, alors que la figure 1.9 représente le noyau total $K_{1/2}$.

Pour illustrer l'efficacité de notre procédure on considère X une variable aléatoire gaussienne tronquée entre -1 et 1, c'est-à-dire si $Y \sim \mathcal{N}(0, \sigma^2)$ alors X n'est autre que la variable Y conditionnée par l'événement $-1 \leq Y \leq 1$. On compare alors sur le bord droit, à fenêtre h fixée (ici $h = 0.6$, ce qui fait que le bord droit n'est autre que l'intervalle $[0.4, 1]$) un estimateur à noyau non modifié (on a pris un noyau d'Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \leq 1\}$) avec notre modification à l'odre 1. Les résultats sont reportés sur la figure 1.10. On constate que l'estimateur modifié corrige la tendance à la sous-estimation de l'estimateur à noyau classique, par contre on récupère un estimateur plus oscillant à cause de la modification polynomiale à l'intérieur.

Pour illustrer numériquement la qualité de notre méthode on répète à fenêtre fixée ($h = 0.6$) l'expérience suivante : partant d'une famille d'observations on compare la

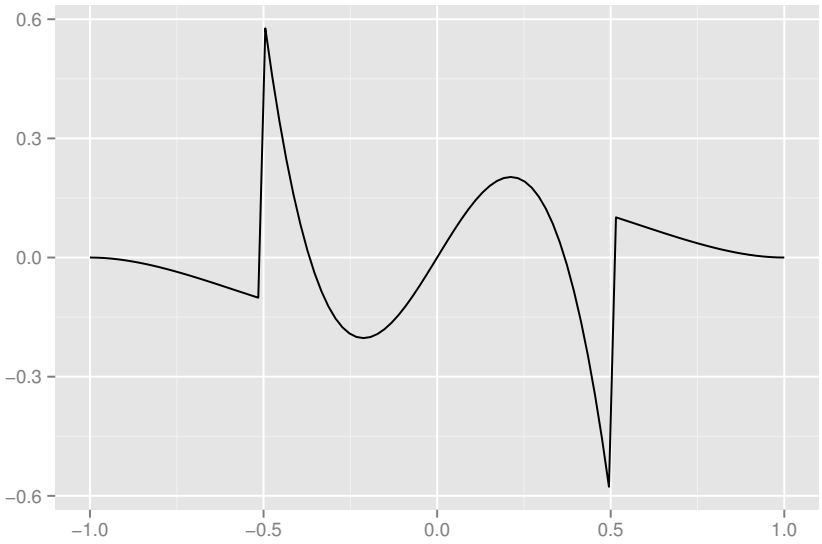

FIGURE 1.8. Terme de correction $\tilde{K}_\alpha$ pour un noyau initial d'ordre 3 pour $\alpha = 0.5$.

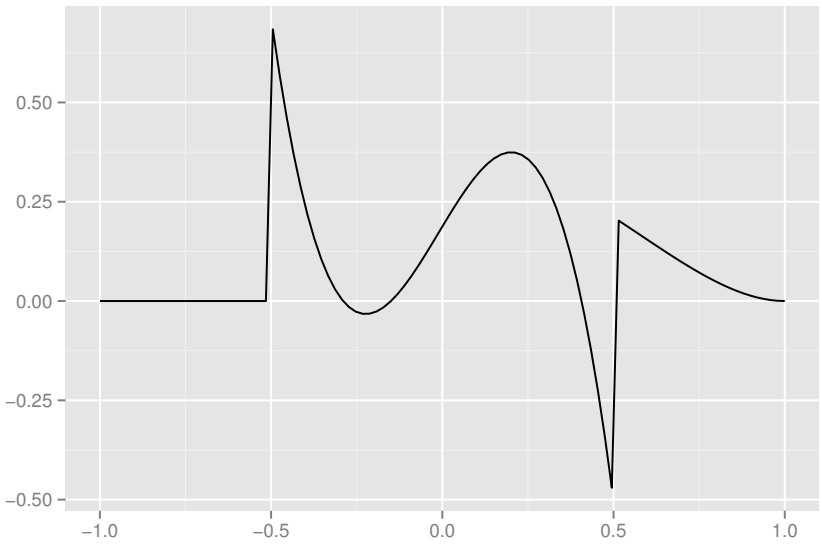

FIGURE 1.9. Noyau corrigé K_α pour un noyau initial d'ordre 3 pour $\alpha = 0.5$.

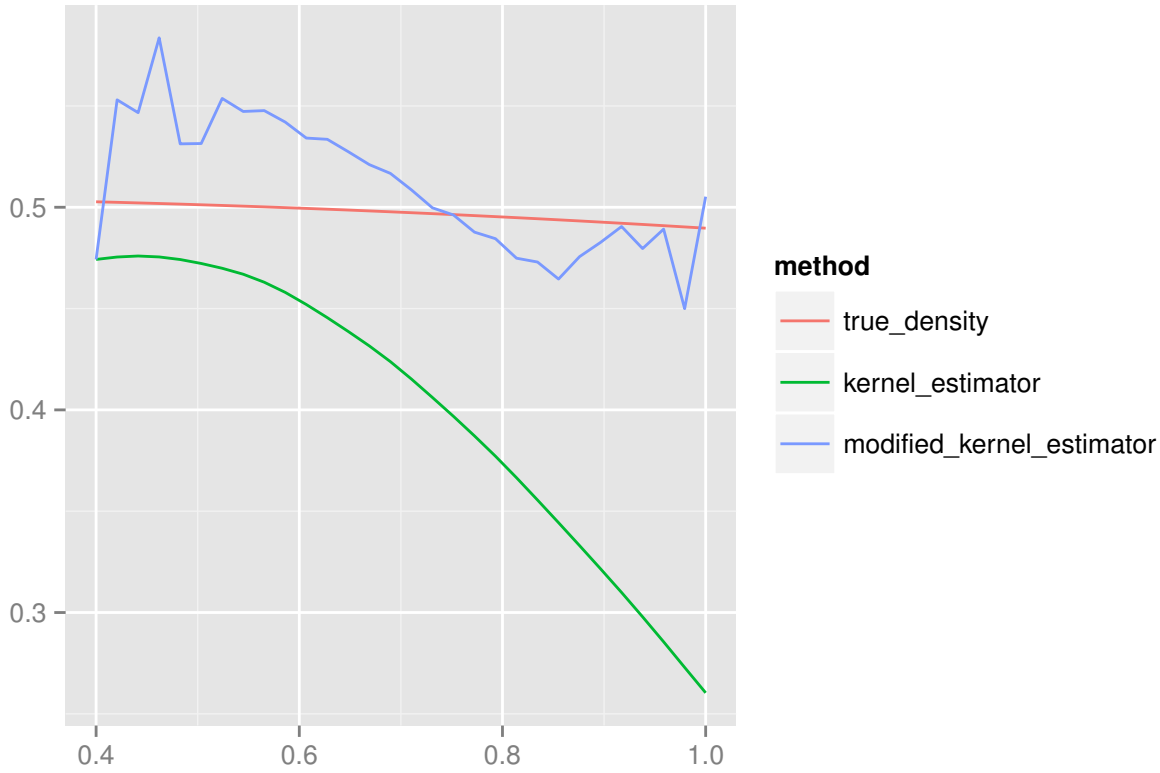

FIGURE 1.10. Estimation de la densité d'une gaussienne tronquée entre -1 et 1, de moyenne nulle et de variance 16, avec un noyau d'Epanechnikov et sa modification au bord pour une fenêtre $h = 0.6$.

probabilité assignée au bord par une méthode à noyau et notre modification (relativement à la vraie probabilité du bord). Les résultats pour 100 répétitions sont reproduits table 1.2. On constate que l'estimateur à noyau sous-estime toujours la probabilité du bord, alors que notre méthode à tendance à la surestimer mais dans des proportions plus de deux fois moindres. Par contre cela s'accompagne d'un accroissement de la variance de la procédure ce qui est normal, étant donné que notre méthode tend à augmenter la norme L^2 du noyau, mais dans des proportions faibles.

	Kernel Estimator	Modified Kernel Estimator
Mean	-18.40 %	7.63 %
Standard Deviation	2.89 %	3.70 %

TABLE 1.2. Comparaison de l'estimateur à noyau et de notre modification (en pourcentage relativement à la vraie densité) pour la gaussienne tronquée de la figure 1.10 et 100 répétitions.

Chapter 2

Orthogonal One Step Greedy Procedure for heteroscedastic linear models

This chapter is the replica of an article submitted to a scientific review. It can be read independently from the rest of the manuscript.

Abstract

This paper investigates the prediction problem in the general Gaussian linear model with correlated noise, under the assumption that the covariance matrix is known, and focuses particularly on the high dimensional setting. We adapt an overly greedy procedure, where the relevant covariates are selected initially in one pass on the data, without any iteration, nor optimization. A simple componentwise regression, followed by an adaptive thresholding, locates leaders among the regressors to reduce the initial dimensionality. A second adaptive thresholding is performed on the linear regression upon the leaders. These steps take into account the correlated structure of the noise, by using weights associated to the covariates in a modified norm induced by the covariance matrix of the noise. The consistency of the procedure is investigated, and rates are provided for a wide range of sparsity classes, with little restriction on the number of regressors. An extensive computational experiment is conducted to emphasize the fact that the good theoretical results are corroborated by quite good practical performances in the presence of correlated noise.

Contents

1 Introduction

Consider the following linear model

$$Z = \Psi\alpha + \eta,$$

where we observe the n-dimensional vector Z, and the $n \times p$ design matrix Ψ. The p-dimensional vector α is the signal to be estimated, while the n-dimensional vector η is an unobservable noise. The case where the number of regressors p is large compared to the number n of observations is the focus of a lot of attention in contemporary statistics. Indeed, such models have many practical applications ranging from genomics, where the number of possibly involved genes in a pathology can be huge compared to the little number of affected people, to image analysis, where the number of unknown pixels can be very large compared to the number of measurements. Natural language processing is another important field of applications: document-term matrices, where each line represents a text from a given corpus and each column a word belonging to one of the texts, leading necessarily to very high dimensional models.

The problem of estimating α in such a high dimensional setting is impossible to solve in full generality. But it can become feasible if some measure of the intrinsic dimension of the signal is in fact much smaller than the dimension of the ambient space $\mathbb{R}^p$. This is referred to as the sparsity of the signal. Many computationally reasonable and theoretically efficient algorithms have been proposed in the literature, using greedy methods [68], [90], [77], [102] or the extraordinary explosive domain of ℓ_1 penalties which we can barely reference: [89], [17], [93] being a few of the references on the topic. For a much more complete bibliography we can refer to [13].

Besides the sparsity of the signal, other conditions appear to be also necessary to solve the problem, basically to prevent multi-colinearities for the columns of the matrix Ψ. Most of the results in the papers cited above are obtained under RIP type-conditions. Recall that the Gram-matrix associated to the subset $\mathcal{C}$ of $\{1, \dots p\}$ is defined by $G(\mathcal{C}) = n^{-1}\, {}^t\Psi_{\mathcal{C}}\Psi_{\mathcal{C}}$ where $\Psi_{\mathcal{C}}$ is the restriction of the matrix Ψ to the columns with indices in $\mathcal{C}$. Roughly speaking the Restricted Identity Property (RIP) means that $G(\mathcal{C})$ is almost the identity matrix as soon as the cardinality $m = |\mathcal{C}|$ is small enough. However this condition can seem quite drastic if the problem is only to avoid too many multi-colinearities. Indeed, one could imagine for instance, replacing '$G(\mathcal{C})$ is almost the identity matrix' by the more flexible condition : $G(\mathcal{C})$ is an invertible matrix. And one might wonder how the results would be affected by such a less restrictive condition. The answer to this question is quite unclear, and one goal of this paper is to shed some light on this aspect.

The problem appears in quite a clear way for instance in models derived from inverse problems where the eigenvalues of the matrices $G(\mathcal{C})$ can depend in a crucial way on the set $\mathcal{C}$. An example of such a case occurs when Ψ is in fact the multiplication of a $n \times n$ symmetric definite positive matrix K by a $n \times p$ matrix X obeying RIP conditions. In practice this is corresponding to a compressed sensing situation where the responses are not only perturbed by noise but are also blurred by the filter K.

An equivalent problem is the heteroscedastic setting, where instead of assuming that the noise components are independent identically distributed random variables, we suppose that the vector η has a covariance matrix Γ. Obviously, if the matrix Ψ satisfies RIP conditions, the transformed model obtained by multiplying by $\Gamma^{-1/2}$ generally no longer satisfies RIP conditions apart from the fact that this 'stabilizing' operation could become rapidly unstable in practice. This paper will focus on this heteroscedastic case.

Although most of the works cited above have been investigating the homoscedastic setting, several works have been conducted in this direction where the noise has a non trivial covariance, studying the behavior of the classical lasso estimator [89], or the adaptive lasso estimator [103] in this correlated setting. In [31], [95] it is proved that the adaptive lasso is consistent and asymptotically normal in a heteroscedastic setting with p fixed

and n growing, but with suboptimal variance. A correction is proposed with a weighted adaptive lasso estimator which has optimal asymptotic variance. In [94] this analysis is extended to the more general bridge estimators. A modification of Lasso and Pseudo-Lasso in the context of linear instrumental variables models able to handle the heteroscedastic setting even with unknown covariance is proposed in [4], where sharp convergence rates are proved under the hypothesis that $\log p = o(n^{1/3})$. In [54] it is shown that the lasso is sign consistent in a Poisson-like model when the signal to noise ratio is large enough.

In the series of papers [63], [73], [74], [72], it has been proved that overly greedy algorithms, which extend Orthogonal Matching Pursuit by incorporating many covariates at their first step, can behave in the high dimensional setting almost as well as much more sophisticated procedures involving optimization steps. The strength of this kind of method is its extreme simplicity. As a drawback, they rely for instance on coherence conditions instead of RIP assumptions. In the context of heteroscedasticity we will see that precisely the simplicity of these types of condition becomes helpful to disentangle with the parts linked to the covariance.

Hence the aim of this paper is to adapt this Orthogonal One-Step Greedy (OOSG) methodology to the case of heteroscedastic high dimensional linear models (even if it is not necessary, the algorithm is still usable for classical low dimensional models, where the number of observations is larger than the number of covariates).

As will be seen, even in this more sophisticated context, the procedure will not require much more complexity in the computations. It will also be quite easy to understand the conditions under which it proves to have theoretical as well as practical good behavior.

Another interesting aspect is that the theoretical conditions under which we can prove the rates of convergence of the algorithm are quite clear extensions of the conditions appearing in inverse models for instance (see [64] for example).

One-Step Greedy procedures are typical selection / estimation procedures in the sense of [47]: in a first step they select a number N of covariates by independent screening [46], then perform least squares regression on those covariates, the resulting estimator being finally thresholded. The number N and the threshold are data driven, giving an adaptive procedure. To adapt to the covariance structure of the errors, we will modify the methodology only by incorporating in the thresholds weights related to the size of the columns of the design in the norm induced by the covariance matrix of the noise.

We show that the rates of OOSG are driven by the standard behavior of an inverse problem term involving additionally the coherence and the sparsity of the signal, together with a term taking into account the location of the signal among the regressors. Indeed, a basic effect of the presence of a non standard covariance is to bring disparity between the potential precisions of estimation of each coordinate of the signal.

Of course, because of the straightforwardness of the method, some loss of efficiency is expected compared to more costly procedures. But even with limitations from both theoretical and practical points of view it performs quite well with small coherence, and is

computationally very attractive. An intensive calculation program has been performed in section 5 to show the advantages and limitations of the method, and it is compared to the weighted adaptive lasso of [31].

2 The Setup

2.1 The model

In this paper, we observe a pair $(Z, \Psi) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$, where Ψ is the (correlated) design matrix and Z a vector of response variables. These two quantities are linked by the standard linear model

$$Z = \Psi\alpha + \eta, \tag{2.1}$$

where the parameter $\alpha \in \mathbb{R}^p$ is the unknown vector to be estimated, and η is a noise perturbation.

We make the following assumptions on model 2.1:

— the vector $\eta = {}^t(\eta_1, \ldots, \eta_n)$ is a (non observed) vector of random errors. It is assumed to be a correlated centered Gaussian vector with distribution $\mathcal{N}_n(0, \Gamma)$, with correlation matrix Γ positive definite and known,

— $\Psi = [\psi_1, \ldots, \psi_p]$ is a $n \times p$ known matrix (this paper focuses mostly on the high-dimensional setting where $p \gg n$, but it is not necessary), where the $\psi_i \in \mathbb{R}^n$ are its column vectors. We assume that Ψ has normalized columns (or normalize them) with respect to the empirical 2-norm, i.e. in the following sense:

$$\frac{1}{n}\|\psi_l\|_2^2 := \frac{1}{n}\sum_{i=1}^{n} \psi_{il}^2 = 1, \quad \forall\, l = 1, \ldots, p. \tag{2.2}$$

2.2 Notation

We define a vector of weights as a vector $w = (w_1, \ldots, w_p) \in \mathbb{R}^p$, such that $w_i \geq 0$ for all $1 \leq i \leq p$. We associate to any $\alpha \in \mathbb{R}^p$ and weight vector w the set:

$$S_w(\alpha) = S_w = \{l;\, |\alpha_l| > w_l\}.$$

In particular, $S_0(\alpha) = S_{(0,\ldots,0)}(\alpha)$ is the support of α.

For $p \in (0, \infty]$, we write $\|.\|_p$ for the usual ℓ_p (pseudo)-norm. For vectors in $\mathbb{R}^p$, define the ℓ_0 quasi-norm as

$$\|x\|_0 = |S_0(\alpha)| = |\{j; x_j \neq 0\}|.$$

We say that a vector of $\mathbb{R}^p$ is K-sparse, $1 \leq K \leq p$, when $\|x\|_0 \leq K$.

Given a $n \times n$ symmetric positive definite matrix Δ, we define the following Δ-norm on $\mathbb{R}^n$:

$$\forall\, x \in \mathbb{R}^n, \quad \|x\|_\Delta^2 =< x, \Delta x >= \|\Delta^{1/2} x\|_2^2,$$

where $< ., . >$ denotes the usual scalar product on $\mathbb{R}^n$. Of particular importance in this paper is the Γ-norm, $\|.\|_\Gamma$, induced by the covariance matrix of the noise η in the model 2.1.

Suppose that T is a subset of $\{1, \ldots, p\}$. We define the restriction of a vector $x \in \mathbb{R}^p$ to the set T as:

$$x_T = \begin{cases} x_i, & \text{if } i \in T \\ 0, & \text{otherwise.} \end{cases}$$

We occasionally abuse this notation and treat x_T as belonging to $\mathbb{R}^T$. In the same way we define the restriction Ψ_T of a $n \times p$ matrix Ψ as the matrix whose columns are listed in the set T. We denote by V_T the subspace of $\mathbb{R}^n$ spanned by the columns of Ψ_T, and by P_{V_T} the orthogonal projection over V_T .

3 The One Step Greedy Algorithm for Heteroscedastic Noise

3.1 Intuition

The principal theoretical intuition behind our procedure comes from the efficiency of thresholding procedures in the "low dimensional" case where $p \leq n$. Indeed if in the model , the design matrix Ψ is one-to-one (so that necessarily $p \leq n$), it admits a left inverse $\Psi^- = ({}^t\Psi\Psi)^{-1\,t}\Psi$. Then $\Psi^- Z = \hat{\alpha}$ is the usual least squares estimator of α, and $\hat{\alpha} \sim \mathcal{N}(\alpha, \Psi^- \Gamma\, {}^t\Psi^-)$, which is the classical Gaussian sequence model with colored noise.

Thresholding in the Gaussian sequence model with colored noise has been studied in [58], [57], [64] extending the white noise results of [38], [42]. It is shown in those papers that a simple pointwise thresholding of $\hat{\alpha}$ is sufficient to provide an optimal estimator (in a minimax sense) for the estimation error. We would like to adapt that procedure to the high dimensional case, where $p > n$, with colored noise, just as [63], [73] extended the low-dimensional white noise results to the high-dimensional case with white noise.

The first obstacle in doing so is that Ψ cannot be one-to-one anymore, and as such does not admit a left inverse. To circumvent this issue we proceed in two steps. First we want to reduce the initial dimensionality of the problem, in the spirit of [46], by selecting only a small (relatively to p) number of informative covariates. For such a selection to be effective, we have to assume some sparsity of the signal of interest α, so that only a small number of covariates are involved in the observed signal Z. Once we have restricted ourselves to use a given subset of covariates, we are facing a "classical" linear model with more observations than covariates and colored noise, to which we can apply the thresholding machinery of

[58], [57], [64]: we can perform the computation of the least squares estimator of the signal in this restricted model, and finally threshold it pointwise.

To perform the initial selection we assume our design Ψ not to be too correlated (this is made precise in section 4.1). If so we can still expect the Gram matrix $\frac{1}{n}{}^t\Psi\Psi$ to behave almost as an isometry on sparse enough signals. Then the vector $\check{\alpha} = \frac{1}{n}{}^t\Psi Z$ would be a good initial proxy of the signal α, from which we base our selection procedure.

Since the noise we consider is heteroscedastic, the variances of the components of $\check{\alpha}$ are not equal. Indeed a simple computation shows that:

$$\text{Var } (\check{\alpha}_l) = \frac{\|\psi_l\|_\Gamma^2}{n^2}.$$

This anisotropic behaviour is the root of all the necessary modifications that we perform in this paper.

3.2 Overview

In this section we describe the estimation of the unknown parameter of interest α using the orthogonal one-step greedy (OOSG) procedure. As input OOSG requires observable data and parameters.

— **Data** the observed response variable Z, the design matrix $\Psi = [\psi_1, \dots, \psi_p]$, and the covariance matrix of the noise, Γ.

— **Parameters:** the initial selection step of the procedure requires a threshold parameter $\lambda_1 \geq 0$ and two size parameters $\mathbf{N}$ and Σ. The final step of the method requires a second threshold parameter $\lambda_2 \geq 0$.

The size parameters have two purposes:

1. we want to control the cardinality of the selected subset of indices with the parameter $\mathbf{N}$, so that our procedure will produce at most $\mathbf{N}$-sparse estimators. We want $\mathbf{N}$ to be small enough to guarantee that the restricted design on any subset of covariates of cardinality bounded by $\mathbf{N}$ is one-to-one, and so that the restricted least squares estimator is well defined (this is quantified precisely in section 4.1).
2. at the same time we want to control the variance supported by the selected set of covariates, that we define, given I a subset of indices, $I \subset \{1, \dots, p\}$, as:

$$\sigma^2_{\text{tot}}(I) = \sum_{l \in I} \text{Var } (\check{\alpha}_l) = \sum_{l \in I} \frac{\|\psi_l\|_\Gamma^2}{n^2}.$$

Then we will constrain our selection step to return a subset of covariates $\mathbf{L}$ such that $\sigma^2(\mathbf{L}) \leq \Sigma$.

It is then convenient to introduce the quantity $\sigma^2(I)$, defined on any subset of indices I as:

$$\sigma^2(I) = \sum_{l \in I} \frac{\|\psi_l\|_\Gamma^2}{n^2} \vee 1 := \sum_{l \in I} \sigma_l^2,$$

and obviously verifies that, for any I, $|I| \leq \sigma^2(I)$ and $\sigma^2_{\text{tot}}(I) \leq \sigma^2(I)$.

The first selection step seeks a subset of covariates, $\mathbf{L}$, based on the relative importance of the indices $1 \leq l \leq p$ in the renormalized vector $|\check{\alpha}|/\sigma$, where σ is the vector $\sigma = (\sigma_1, \ldots, \sigma_p)$ (and division is understood componentwise). To avoid including in the leader set covariates with very low predictive power, we only select covariates whose normalized importance $|\check{\alpha}_l|/\sigma_l$ is larger than some provided threshold λ_1. On the other hand we restrict the size of the leader set by requiring that $\sigma^2(\mathbf{L}) \leq \Sigma$, and $|\mathbf{L}| \leq \mathbf{N}$, where Σ and $\mathbf{N}$ are critical size parameters provided by the user,.

Once this set is selected we compute the restricted least squares estimator of α on the sub-model of the leading covariates. Finally we threshold adaptively the resulting estimator, still taking the heteroscedastic setting into account by using the weight vector σ, at another provided level λ_2.

In a more precise fashion we can summarize our procedure as follows:

1. we select one-by-one in our leader set $\mathbf{L}$ covariates whose indices belong to

$$S_{\lambda_1 \sigma}(\check{\alpha}) = \{1 \leq l \leq p, |\check{\alpha}_l/\sigma_l| \geq \lambda_1\}$$

   from the largest ratio $|\check{\alpha}_l/\sigma_l|$ to the lowest. Each time we incorporate a coordinate l in $\mathbf{L}$ we check that $\sigma^2(\mathbf{L}) \leq \Sigma$ until we have selected the whole set $S_{\lambda_1 \sigma}(\check{\alpha})$, that we have select $\mathbf{N}$ covariates, or that the total weight of incorporated coordinates would exceed Σ by incorporating a new one, in which case we stop. Ties are broken lexicographically.
2. having selected the set $\mathbf{L}$ we construct an estimator $\hat{\alpha}(\mathbf{L})$ by restricted least squares: $\hat{\alpha}(\mathbf{L}) = \left({}^t\Psi_{\mathbf{L}}\Psi_{\mathbf{L}}\right)^{-1} {}^t\Psi_{\mathbf{L}} Z$.
3. finally we adaptively threshold the resulting estimator $\hat{\alpha}(\mathbf{L})$ at level λ_2:

$$\hat{\alpha}(\mathbf{L}, \lambda_2) = \begin{cases} \hat{\alpha}(\mathbf{L})_l & \text{if } |\hat{\alpha}(\mathbf{L})_l/\sigma_l| \geq \lambda_2, \\ 0 & \text{if } |\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2. \end{cases}$$

It is important to underline the computational efficiency of such a procedure. Indeed it can be seen as one step of orthogonal matching pursuit, where we let all the relevant covariates enter the set of selected atoms in this initial step. No iteration is required.

3.3 Pseudocode description of the method

Details of the procedure are described in the following pseudocode. It can be noticed already that the procedure requires no optimization nor iteration, and so is very computationally efficient.

Input: observed data Z, design Ψ, covariance Γ, tuning parameters λ_1, λ_2, Σ, $\mathbf{N}$
Output: estimated parameter $\hat{\alpha}(\mathbf{L}, \lambda_2)$, and predicted response $\hat{Z}$

for i = 1:p **do**
  $\sigma_l^2 \leftarrow 1 \vee \frac{\|\psi_l\|_\Gamma^2}{n}$
end for

for i = 1:p **do**
  $\check{\alpha}_i \leftarrow \frac{1}{n} < \psi_i, Z >$ ▷ Componentwise regression
end for

$S \leftarrow \{i; |\check{\alpha}_i/\sigma_i| \geq \lambda_1\}$
$\mathbf{L} \leftarrow \emptyset$
$w \leftarrow 0$ ▷ Weight counter

while $w \leq \Sigma$ and $|\mathbf{L}| \leq \mathbf{N} - 1$ **do** ▷ Selecting the set of leaders
  $k_m \leftarrow which.max(\{|\check{\alpha}_i/\sigma_i|, i \in \check{S}_{\lambda_1}\})$
  if $w + \sigma_{k_m}^2 \leq \Sigma$ **then**
    $\mathbf{L} \leftarrow \mathbf{L} \cup \{k_m\}$
    $\check{S}_{\lambda_1} \leftarrow \check{S}_{\lambda_1} \setminus \{k_m\}$
    $w \leftarrow w + \sigma_{k_m}^2$
  else
    Break
  end if
end while

$\hat{\alpha}(\mathbf{L}) \leftarrow ({}^t\Psi_{\mathbf{L}}\Psi_{\mathbf{L}})^{-1}\,{}^t\Psi_{\mathbf{L}}Z$ ▷ Restricted least squares estimator
$\mathbf{L}^+ \leftarrow \{l \in \mathbf{L}, |\hat{\alpha}(\mathbf{L})_l|/\sigma_l \geq \lambda_2\}$
for $i \in \mathbf{L}$ **do** ▷ Thresholding
  if $i \in \mathbf{L}^+$ **then**
    $\hat{\alpha}(\mathbf{L}, \lambda_2)_i \leftarrow \hat{\alpha}(\mathbf{L})_i$
  else
    $\hat{\alpha}(\mathbf{L}, \lambda_2)_i \leftarrow 0$
  end if
end for

$\hat{Z} \leftarrow \Psi\hat{\alpha}(\mathbf{L}, \lambda_2)$

4 Theoretical Results

4.1 Coherence

This section aims at quantifying the notion of correlated design. Given Ψ a $n \times p$ design, and $\nu \in (0,1)$, define $\mathbf{N}_{\max} = \mathbf{N}_{\max}(\Psi, \nu)$ as the maximum of the positive integers $k > 0$ such that:

$$\forall\, x \in \mathbb{R}^p, \|x\|_0 \leq k \Rightarrow (1-\nu)\|x\|_2^2 \leq \frac{1}{n}\|\Psi x\|_2^2 \leq (1+\nu)\|x\|_2^2. \tag{2.3}$$

This is a slight reformulation of the restricted isometry property of [20]. Thanks to the normalization condition 2.2, we have that $\mathbf{N}_{\max} \geq 1$, for all $\nu \in (0,1)$. It should be noticed already that if T is a subset of indices of cardinality less than $\mathbf{N}_{\max}$, then the Gram matrix $\frac{1}{n}{}^t\Psi_T\Psi_T$ is invertible since its eigenvalues are contained in $[1-\nu, 1+\nu]$.

Using eq. (2.3) and the variational characterization of eigenvalues ([53]) we get the following lemma.

Lemma 2.1. *Let Ψ be a $n \times p$ matrix verifying condition 2.2. Let T be a subset of indices such that $|T| \leq \mathbf{N}_{\max}$. Then for all $x \in \mathbb{R}^{|T|}$:*

1. $\sqrt{1-\nu}\|x\|_2 \leq \|\frac{1}{\sqrt{n}}\Psi_T x\|_2 \leq \sqrt{1+\nu}\|x\|_2$,
2. $(1-\nu)\|x\|_2 \leq \|\frac{1}{n}{}^t\Psi_T\Psi_T x\|_2 \leq (1+\nu)\|x\|_2$,
3. $\frac{1}{(1+\nu)}\|x\|_2 \leq \|(\frac{1}{n}{}^t\Psi_T\Psi_T)^{-1}x\|_2 \leq \frac{1}{(1-\nu)}\|x\|_2$,
4. *If $x \in \mathbb{R}^p$:*
$$\frac{1}{1+\nu}\frac{1}{n}\|{}^t\Psi_T x\|_2^2 \leq \|P_{V_T}x\|_2^2 \leq \frac{1}{1-\nu}\frac{1}{n}\|{}^t\Psi_T x\|_2^2.$$

This is proved, for the sake of completeness, in section 7.1. The important take away from lemma 2.1 is that the design Ψ behaves almost as an isometry on sparse enough vectors, or equivalently that the Gram matrix $\frac{1}{n}{}^t\Psi\Psi$ is close to the identity operator if restricted to $\mathbf{N}_{\max}$-sparse vectors.

Usually finding the quantity $\mathbf{N}_{\max}$ given the design Ψ is computationally intractable, it is therefore common practice ([29],[36],[2]) to quantify more crudely the internal correlation among the predictors using the coherence of the design: the largest, in absolute value, inner product of two different columns of the design. Formally we define the coherence of the design Ψ as:

$$\tau_n := \max_{l \neq m} \frac{1}{n}|<\psi_l, \psi_m>|. \tag{2.4}$$

The coherence is a crude, but computable, measure of the correlation among the covariates of the design. In the same way we can define a coherence relative to the scalar product induced by the covariance Γ of the noise as:

$$\tau_n(\Gamma) = \max_{l \neq l'} \frac{|<\psi_l, \Gamma\psi_{l'}>|}{\|\psi_l\|_\Gamma \|\psi_{l'}\|_\Gamma}.$$

The coherence of a design allows to lower bound the quantity $\mathbf{N}_{\max}$, as stated in the following lemma.

Lemma 2.2. *Let Ψ be a $n \times p$ design with normalization condition 2.2. Let $\nu \in (0,1)$ and let τ_n be the coherence of Ψ, then:*

$$\mathbf{N}_{\max} = \mathbf{N}_{\max}(\Psi, \nu) \geq \lfloor \nu/\tau_n \rfloor + 1 \tag{2.5}$$

For a proof we refer to section 7.2. As a consequence, the coherence τ_n provides us with a bound on the sparsity of signals on which the normalized design $\frac{1}{\sqrt{n}}\Psi$ acts almost as an isometry.

Furthermore the coherence measures how well the initial proxy of the signal, $\check{\alpha} = \frac{1}{n}{}^t\Psi Z$, will approximate the signal α. Simple computations are summarized in the following lemma.

Lemma 2.3. *Let $\check{\alpha} = \frac{1}{n}{}^t\Psi Z$ be our initial proxy of the signal α. Then for all indices $1 \leq l \leq p$:*

$$\check{\alpha}_l = \alpha_l + R_l + \eta_l,$$

where $R_l = \frac{1}{n}\sum_{i \neq l} \alpha_i < \psi_i, \psi_l >$, and $\eta_l = \frac{1}{n} < \eta, \psi_l >$. The noise η_l is normally distributed, $\eta_l \sim \mathcal{N}\Big(0, \frac{\|\psi_l\|_\Gamma^2}{n^2}\Big)$, and for all l, $|R_l| \leq \tau_n \|\alpha\|_1$.

As the consequence the lower the coherence, the less biased is the approximation $\check{\alpha}$.

4.2 Rates of convergence of OOSG on weighted ℓ_q balls

We define the prediction error of our procedure as:

$$\mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha})\|_2^2\Big].$$

The next theorem precises the effectiveness of the method under general parameter settings, making explicit the importance of the coherence, for the prediction error.

Theorem 2.4. *Let $K_1, K_2, \theta > 0$ be constants. Suppose that for $\lambda_1 > 0$, $|S_{\frac{\lambda_1}{2}\sigma}(\alpha)| \leq \mathbf{N}_{\max}$. If we choose the parameters of the procedure such that:*

1. $\tau_n N \leq K_1$ *and* $\tau_n(\Gamma)\Sigma \leq K_2$,
2. $\Sigma \leq p^\theta$, $\mathbf{N} \leq \mathbf{N}_{\max}$,
3. $\lambda_1^2 \geq C_2\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$,
4. $\lambda_2 \leq \lambda_1$,
5. $\Sigma \geq \sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha))$,
6. $\mathbf{N} \geq |S_{\frac{\lambda_1}{2}\sigma}(\alpha)|$,

then, if $p \geq 2$, there exists a constant C depending on ν, K_1, K_2, θ such that

$$\mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha}(\boldsymbol{L}, \lambda_2))\|_2^2\Big] \leq C\Big\{\|\alpha_{S_1^c}\|_2^2 + \Big(\frac{\tau_n}{\nu} \wedge 1\Big)\|\alpha_{S_1^c}\|_1^2 \Big(\tau_n^2\|\alpha\|_1^2 + \sigma_{\max}^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n} + \lambda_1^2\Big)\sigma^2\Big(S_2\Big)\Big\},$$

where $S_1 = S_{2\lambda_1\sigma}(\alpha)$ and $S_2 = S_{\frac{\lambda_2}{2}\sigma}(\alpha)$, and for all subset of indices S,
$\sigma_{\max}^2(S) = \max_{l \in S} \frac{\|\psi_l\|_\Gamma^2}{n} \vee 1$.

Consider signals α satisfying a weighted ℓ_q ball sparsity constraint defined as follow:

— for $q \in (0,1]$, $\mathcal{B}_{q,\sigma}(M) = \left\{\alpha \in \mathbb{R}^p; \left(\sum_{l=1}^p \sigma_l^2 |\alpha_l/\sigma_l|^q\right)^{1/q} \leq M\right\}$,

— for $q = 0$, $\mathcal{B}_{0,\sigma}(S,M) = \left\{\alpha \in \mathbb{R}^p; \sum_{l=1}^p \sigma_l^2 1\{\alpha_l \neq 0\} \leq S, \|\alpha\|_1 \leq M\right\}$.

When specialized to those signals, and with an extra assumption on the coherence, we get rates of convergence of our procedure on those weighted ℓ_q ball.

Theorem 2.5. *Suppose that $\tau_n \leq c\sqrt{\frac{\log p}{n}}$, for some constant $c > 0$. Then we set the parameters of our method in the following manner:*

— *let $\mathbf{N} = \frac{\nu}{\tau_n} \vee 1$,*

— *let $p^\theta \geq \Sigma \geq \frac{\nu}{\tau_n} \vee 1$, for some constant $\theta > 0$, and $\tau_n(\Gamma)\Sigma \leq K$, $K > 0$,*

— *set the threshold $\lambda_1^2 \geq C_1 \frac{\log p}{n}$, where the constant C_1 depends on M, and $\lambda_2^2 \geq C_2 \frac{\log p}{n}$, while $\lambda_2 \leq \lambda_1$.*

Under this setting:

1. *there exists a constant $C > 0$, depending on ν, C_1, C_2, K, θ and c such that:*

$$\forall \alpha \in \mathcal{B}_{q,\sigma}(M), \quad \mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha}(\boldsymbol{L}, \lambda_2))\|_2^2\Big] \leq C_1\, \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\Big(\frac{\log p}{n}\Big)^{1-q/2}.$$

2. *if $S \leq \nu/\tau_n \vee 1$ there exists a constant C, depending on ν, C_1, C_2, K, θ and c such that:*

$$\forall \alpha \in \mathcal{B}_{0,\sigma}(S,M), \quad \mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha}(\boldsymbol{L}, \lambda_2))\|_2^2\Big] \leq C_2\, \sigma^2_{\max}(S_{\lambda_1/2,\sigma}(\alpha))\Big(\frac{S\log p}{n}\Big).$$

For a proof we refer to section 6.5. In fact theorem 2.5 will be proved as a corollary of theorem 2.4.

4.3 Discussion

In [81] it is proved that the minimax error in ℓ_2-prediction loss in a homoscedastic high dimensional linear model scales as $\left(\frac{\log p}{n}\right)^{1-q/2}$ under the constraint that the signal belongs to some ℓ_q ball, $q \in (0,1]$, and $\frac{S\log p}{n}$ if the signal belongs to a ℓ_0 ball (and under regularity conditions on the design). Here we find in the rates the same term $\left(\frac{\log p}{n}\right)^{1-q/2}$ under a weighted ℓ_q ball constraint, $q \in (0,1]$, and $\frac{S\log p}{n}$ under a ℓ_0 constraint (but with an additional constraint, we need a bound on the ℓ_1 norm of the signal). Notice that this weighting is appearing in the same way in inverse problems (see for instance [64]).

The heteroscedastic setting additionally impacts the rates through the term $\sigma^2_{\max}(S_{\lambda_1/2,\sigma}(\alpha))$. This term takes into account the location of the signal which is relevant here since the noise is not isotropic. At this stage, we do not know whether this term is due to our calculation or unavoidably due to the procedure. The main difficulty in this setting compared to LOL coming from the anisotropy of the noise. Indeed the expectation of the norm of the projected noise on a given subspace now does not only depend on the dimension of this subspace as in LOL, but depends too on the particular subspace we choose. This

is reflected in proposition 2.8, and requires, at least in theory, the additional constraint $\sigma^2(\mathbf{L}) \leq \Sigma$.

It is interesting also to notice that when the 'normalizing factors' $\frac{\|\psi_l\|_\Gamma^2}{n}$ are small (less than 1 for instance), the rates are those obtained by the procedure LOL with smaller constants and for larger sets of coefficients.

Comparing for instance to [64], in the case where the design Ψ is orthogonal, and the covariance matrix Γ is diagonalizable in the basis Ψ, then $\tau_n = \tau_n(\Gamma) = 0$, and the term $\sigma^2_{\max}(T_{\lambda_{1/2},\sigma}(\alpha))$ is not present in the rates. But in this case our modified OOSG procedure is equivalent to the usual heteroscedastic hard thresholding, and will perform with optimal rates.

It is interesting also to compare the behaviour of OOSG with the behaviour of the homoscedastic LOL procedure applied to the stabilized model (transformed so that its covariance is the identity). Indeed if we multiply model by $\Gamma^{-1/2}$ we get the model:

$$\Gamma^{-1/2}Z = \Gamma^{-1/2}\Psi\alpha + \Gamma^{-1/2}\eta, \tag{2.6}$$

with $\Gamma^{-1/2}\eta = \tilde{\eta} \sim \mathcal{N}(0, I_n)$. Let $\tilde{\Psi} = \Gamma^{-1/2}\Psi = [\tilde{\psi_1} \dots \tilde{\psi_p}]$. But now $\tilde{\Psi}$ has to be normalized and let D be the normalization diagonal matrix with diagonal coefficients $\frac{\sqrt{n}}{\|\tilde{\psi}_i\|_2}$. Then $\Phi = \tilde{\Psi}D$ verifies the normalization condition 2.2, and if we let $\tilde{\alpha} = D^{-1}\alpha$, the transformed model 2.6 is equivalent to the normalized homoscedastic model:

$$Y = \Phi\tilde{\alpha} + \tilde{\eta}. \tag{2.7}$$

It should be noticed that:

— the behaviour of LOL is governed by the coherence of the new design Φ, i.e. by the Γ^{-1}-coherence of the design Ψ:

$$\max_{i \neq j} \frac{| < \psi_i, \Gamma^{-1}\psi_j > |}{\|\psi_i\|_{\Gamma^{-1}}\|\psi_j\|_{\Gamma^{-1}}},$$

which is a "global" condition. On the other hand the behaviour of OOSG depends on $\sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))$ which is a local quantity, depending on the location of the support of the signal.

To investigate how different these two quantities can be, consider u and v two unitary eigenvectors of Γ associated respectively to two different eigenvalues of Γ, $\lambda, \mu > 0$. Then necessarily u and v are orthogonal. Consider now $w = \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v$, then we suppose that w and v are columns of the design Ψ. Then obviously the coherence of this design is at least $\frac{1}{\sqrt{2}}$. Furthermore we can compute that

$$\frac{| < w, \Gamma^{-1}v > |}{\|w\|_{\Gamma^{-1}}\|v\|_{\Gamma^{-1}}} = \frac{\sqrt{\lambda}}{\sqrt{\mu + \lambda}}.$$

As a consequence if $\mu \ll \lambda$, the Γ^{-1}-coherence is very close to 1 and LOL applied to the transformed model will behave very badly. Indeed the transformed design

will be so correlated that the initial estimator $\check{\alpha}$ will be very highly biased. On the other hand if the signal of interest α is 1-sparse and supported on the covariate v then the behaviour of our modified OOSG procedure is driven by $\|v\|_\Gamma^2 = \mu$, which is much more favorable.

— LOL in the transformed model does not estimate the signal α, but the rescaled one $\tilde{\alpha}$. To estimate α we need to back-transform the estimator obtained by LOL in the transformed model. Indeed let $\hat{\alpha}_{\text{LOL}}$ be the LOL estimator in this transformed model, and $\tilde{\alpha}_{\text{LOL}} = D\hat{\alpha}_{\text{LOL}}$ its back transformation. This back-transformation operation is very sensitive to the knowledge of the whole matrix Γ, and whole procedure uses the matrix Γ twice. It can then be expected to be much more sensitive to errors in the estimation of Γ in practical setting (where Γ usually has to be estimated from the data) than the modified OOSG procedure, which uses the knowledge of Γ only once.

— finally, from a computational point of view, especially in the case where n is very large, the computation of $\Gamma^{-1/2}$ can be intractable (it is generally a very expensive computation). In that case the modified OOSG procedure appears as much more efficient since it does not rely on suh an inversion.

5 Numerical Study

This section is an extensive computational study of the Orthogonal One Step Greedy (OOSG) procedure for heteroscedastic linear models. The performance of OOSG is studied over various ranges of indeterminacy $\delta = 1 - n/p$, sparsity rates $\rho = S/n$ and various covariance structures. Our procedure is then compared to the procedure of [31].

5.1 Experimental Design

The design matrices Ψ considered in this study are of random type and built on $n \times p$ independent and identically distributed normal random variables, a favorable setting for our method since they usually have low coherence. Given Ψ, the target observations are $Z = \Psi\alpha + \eta$, where $\eta \sim \mathcal{N}(0, \Gamma)$. The vector of parameters α is simulated as follows: all coordinates are zero except S picked uniformly at random among the p possible choices. If the l-coordinate has been picked, we take $\alpha_l = r_l|z_l|$, where r_l is a Rademacher random variable, i.e. takes the value $+1$ or -1 with equal probability, and $z_l \sim \mathcal{N}(2, 1)$. The vector α is then renormalized to fix its norm $\|\alpha\|_2^2 = e_s$ to be able to modify the signal to noise ratio.

The covariance matrices Γ will be of two types:

— autocorrelated, usual in time series analysis. To simulate this case we will consider the covariance matrices $\Gamma_{ij}^a = a^{|i-j|}$, for $a < 1$. The larger a is, the more autocorrelated is the noise.

— heteroscedastic, where the variance of an observation is a function of its mean. To simulate this case we will, as in [31], consider matrices Γ_σ which are diagonal and

such that $\Gamma_{ii,\sigma}^{1/2} = \sigma(\psi^i, \alpha)$ where ψ^i are the line vectors of the design and σ a function. We will consider the functions $\sigma(\psi^i, \alpha) = | < \psi^i, \alpha > |$ and $\sigma(\psi^i, \alpha) = \mathrm{e}^{|<\psi^i, \alpha>|}$.
To fix the expectation of the noise energy, $\mathbb{E}[\|\eta\|_2^2] = \mathrm{tr}(\Gamma)$, we re-normalize Γ by $e_n * \frac{1}{\mathrm{tr}(\Gamma)}$ to get a covariance matrix Γ with mean noise energy a given $e_n = \mathrm{Trace}(\Gamma) > 0$.

To evaluate the quality of OOSG we will compute the normalized prediction error $E = \|\Psi\alpha - \hat{Z}\|_2^2 / \|\Psi\alpha\|_2^2$.

5.2 Algorithm

OOSG depends on two thresholding parameters λ_1 and λ_2 and two size parameters Σ and $\mathbf{N}$. The theoretical optimal values of the thresholds are not easy to access in practice, since they depend on parameters (like a bound on the q-norm of the signal) that we cannot evaluate without a priori information. And the theoretical value of $\mathbf{N}$ as the inverse of the coherence is usually too restrictive in practice. We will now explain how we proceed to avoid these difficulties and provide a data driven method.

1. Given Z, Ψ and Γ we start by computing for all the indices $1 \leq l \leq p$ the absolute normalized correlations $c_l = | < \psi_l, Z > | / \|\psi_l\|_\Gamma$.
2. The quantities c_l induce an order on the columns of the design Ψ (ties are broken lexicographically). We start with an empty set of leaders $\mathbf{L}$ and we incorporate to $\mathbf{L}$ indices l in the order induced by the c_l from the greatest to the lowest (this is a completely forward procedure). At each step, i.e. each time we introduce a new index in $\mathbf{L}$, we compute the lowest and greatest singular values $\lambda_{\min}$ and $\lambda_{\max}$ of $\Psi_{\mathbf{L}}$. Now given a parameter $\nu \in (0, 1)$ we stop adding new covariates as soon as $\lambda_{\min} < 1 - \nu$ or $\lambda_{\max} > 1 + \nu$. For each ν on a regular grid we compute the cardinality of the resulting leader set. Obviously while ν decreases this cardinality increases. We stop to decrease ν when the increase in cardinality of the leader set starts to become less than the last time we decreased the parameter. This provides us with a well defined leader set.
3. Having selected the set of leaders $\mathbf{L}$ at the preceding step, we compute the restricted least squares estimator $\hat{\alpha}(\mathbf{L}) = ({}^t\Psi_{\mathbf{L}}\Psi_{\mathbf{L}})^{-1t}\Psi_{\mathbf{L}}Z$.
4. Now we compute, for every index $l \in \mathbf{L}$, the quantities $d_l = |\hat{\alpha}(\mathbf{L})_l| / \|\psi_l\|_\Gamma$. To select the subset $\mathbf{L}^+ \subset \mathbf{L}$ we incorporate in $\mathbf{L}^+$ only indices $l \in \mathbf{L}$ such that $d_l \geq \lambda_2$. The threshold λ_2 is chosen adaptively using a cross-validation criterion.

5.3 Effect of indeterminacy and sparsity ratio

We consider design matrices Ψ whose terms are i.i.d. Gaussian variables $\mathcal{N}(0, 1)$, renormalized such that their columns have unit 2-norm. This is a particularly important case since they are known to have good theoretical properties in terms of restricted isometry

property [20]. As we show on eq. (3.3) they have good coherence properties. We show on eq. (3.3) the mean empirical coherence of these matrices as a function of the indeterminacy for different values p with $K = 200$ repetitions. We see that the empirical coherence is usually very small when δ is small enough, and even smaller for large values of p.

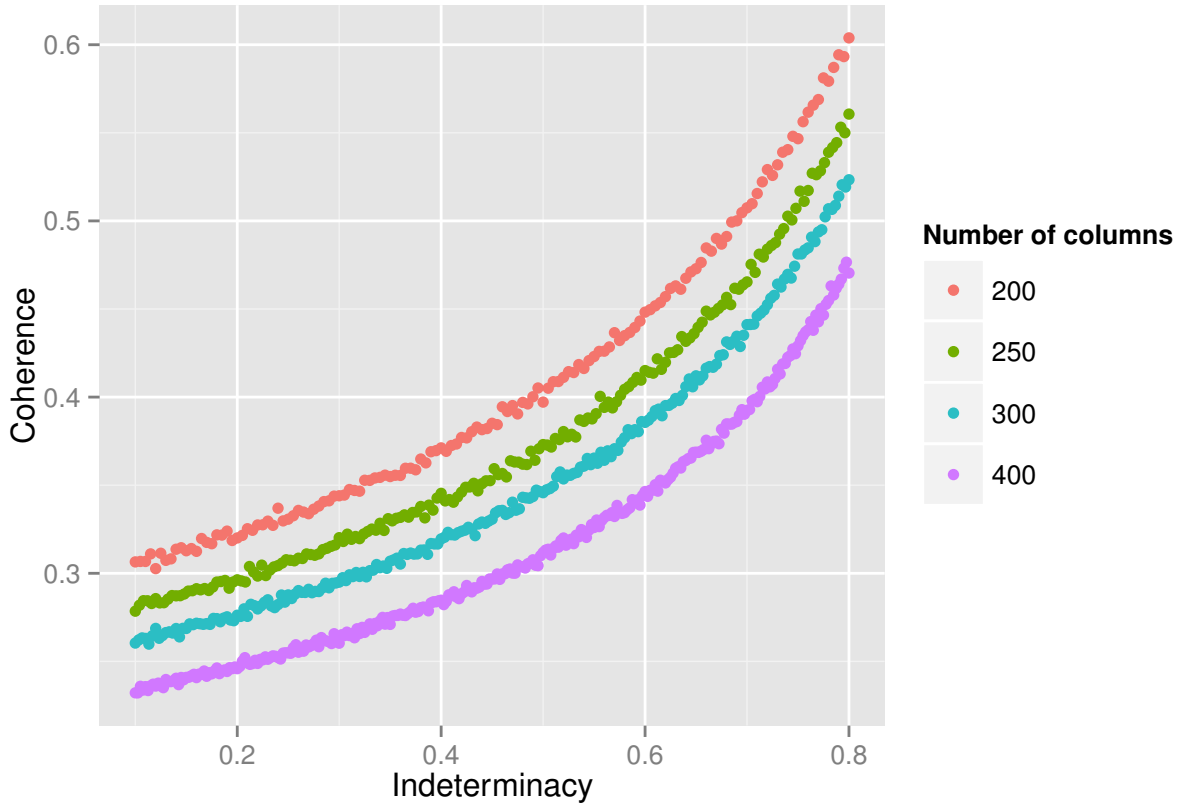

FIGURE 2.1. Empirical mean coherence of a normalized Gaussian random design for different values of p as a function of the indeterminacy, deduced from 200 trials.

The next two experiments aim at illustrating the effects of the indeterminacy level and the sparsity rate on the effectiveness of the OOSG heteroscedastic procedure. Each point on each plot is the mean of the repetition of $K = 200$ trials. Since we do not focus at the moment on the effect of the noise structure we will consider correlated linear models with covariance Γ^a, $a = 0.3$ (autocorrelation structure) renormalized so that the SNR will be fixed at 5.

Influence of the indeterminacy level: fig. 2.2 illustrates the performance of OOSG when the indeterminacy is varying (p is fixed, $p = 150$, and n is varying between 50 and 120) for different sparsity values of the signal. The normalized prediction error E grows with the indeterminacy δ in a linear way. The less sparse the signal, the bigger the error, with a clear jump when the sparsity gets too large ($S = 40$). We see that for low indeterminacy (≤ 30 %) the method performs very well as long as the sparsity is low. It can be noticed that the lower the indeterminacy level, the lower the coherence of the design.

Influence of the sparsity rate: fig. 2.3 studies the performance of OOSG as a

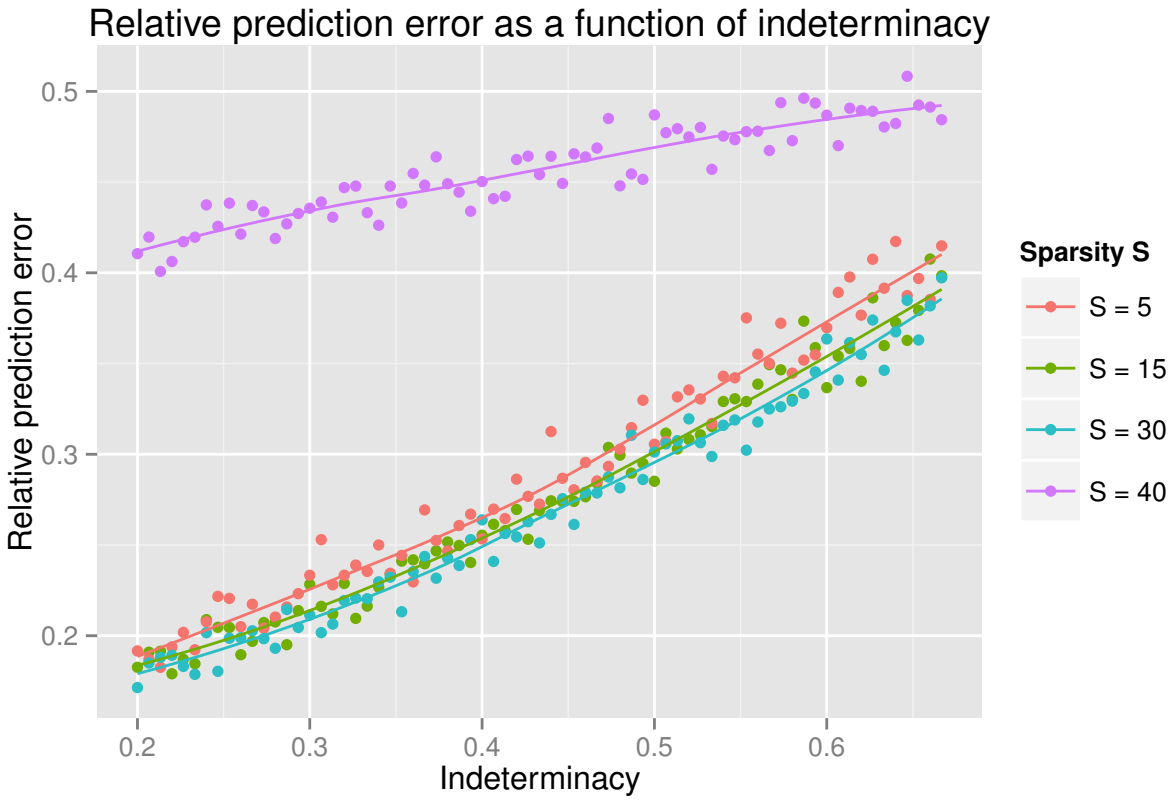

FIGURE 2.2. Normalized prediction error of the HLOL procedure with autocorrelated noise, with covariance Γ^a, $a = 0.3$, as a function of the indeterminacy $\delta = 1 - n/p$ deduced from 200 trials for each point, for different values of the sparsity S. SNR = 5.

function of the sparsity ratio $\rho = S/n$ for different values of indeterminacy. For each curve p and n are kept the same while the sparsity S varies. From a curve to another p is constant and n varies (p is fixed at 150, S varies between 1 and 100). We can see that the performance of the method are very good when the sparsity rate is small, at least as long as the indeterminacy is not too high. When indeterminacy is too high (0.8 here) the method performs poorly. For sparsity ratio ≥ 20 % we see that the method under performs too.

5.4 Comparison with LOL

To illustrate how can OOSG compare to LOL and to LOL applied to the transformed model (WLOL) 2.7:

$$Y = \Phi\tilde{\alpha} + \tilde{\eta}$$

we consider a Gaussian design with a covariance Γ^a, for $a = 0.99$. In that case the covariance matrix is almost singular, and the coherence of ${}^t\Psi\Gamma\Psi$ is much smaller than the coherence of ${}^t\Psi\Gamma^{-1}\Psi$. Accordingly to the theoretical results this is a favorable situation for the HLOL procedure as can be shown on fig. 2.4.

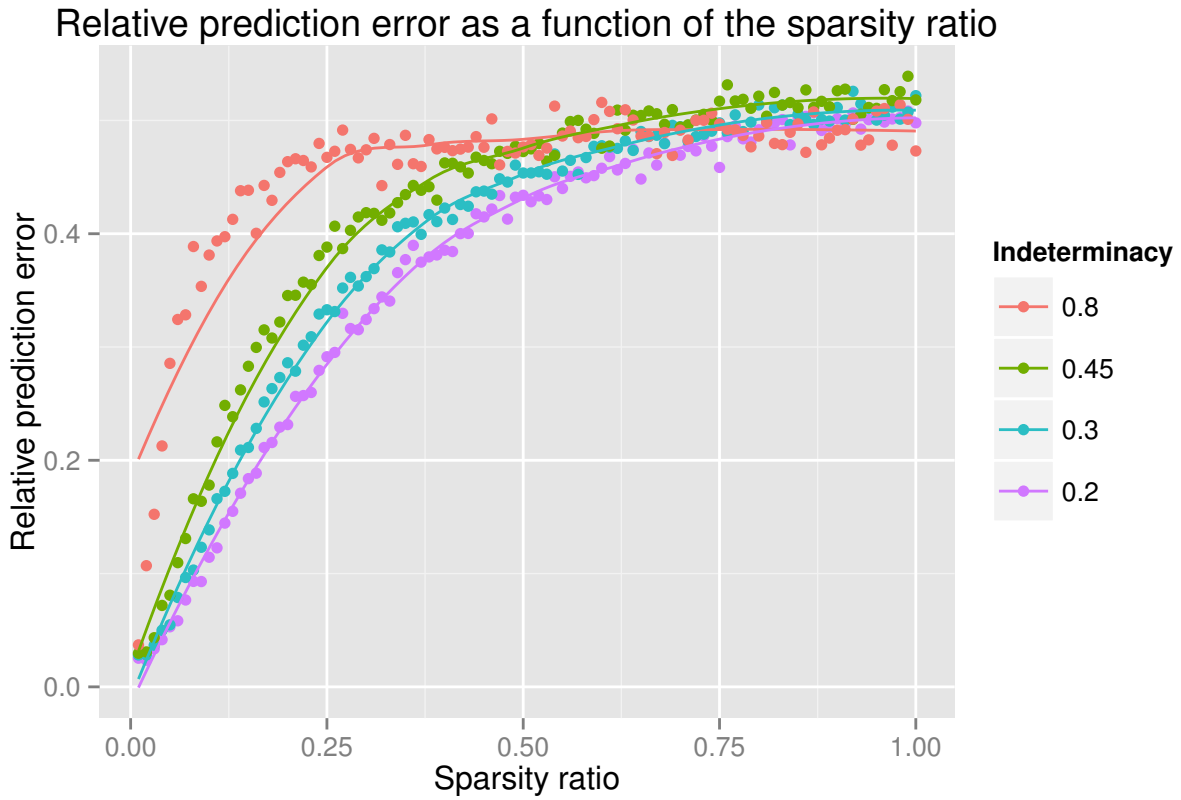

FIGURE 2.3. Normalized prediction error of the HLOL procedure with autocorrelated noise, with covariance Γ^a, $a = 0.3$, as a function of the sparsity rate $\rho = S/n$ deduced from 200 trials for each point, for different values of the indeterminacy δ. SNR $= 5$

5.5 Comparison with weighted adaptive Lasso

In this section we compare the OOSG methodology with the weighted adaptive Lasso estimator of [31]. We consider diagonal covariance matrices of the form $\Gamma^{1/2}_{ii,\sigma} = \sigma(< \psi^i, \alpha >)$, with linear and exponential behavior. Only the function σ is supposed to be known (and not the full matrix Γ since it depends on the signal that we try to estimate). In [31] it is advocated to use the estimator:

$$\hat{\alpha}^\lambda = \text{argmin}_\alpha \Big[\sum_{i=1}^{n} \Big(\frac{Z_i - < \psi^i, \alpha >}{\sigma(\psi^i, \bar{\alpha})} \Big)^2 + \lambda \Big(\sum_{j=1}^{p} \frac{|\alpha_j|}{|\bar{\alpha}_j|} \Big) \Big]$$

where $\bar{\alpha}$ is a preliminary estimator of the signal α. In the next simulations we will take a classical Lasso estimator as a preliminary estimator, and the parameter λ is optimized using cross-validation. To adapt OOSG to this setting we will use a standard LOL methodology to provide us with a preliminary estimator of the signal, which we will then use to get an estimator $\hat{\Gamma}$ of the covariance matrix, that we will be using in the OOSG procedure. We compare OOSG to this estimator in two cases: when the variance function is linear fig. 2.5, and when it is exponential fig. 2.6. In both cases we compare the normalized prediction error of the two method as a function of the sparsity ratio. In the linear case for low sparsity ratios we see that OOSG outperforms weighted adaptive Lasso at least

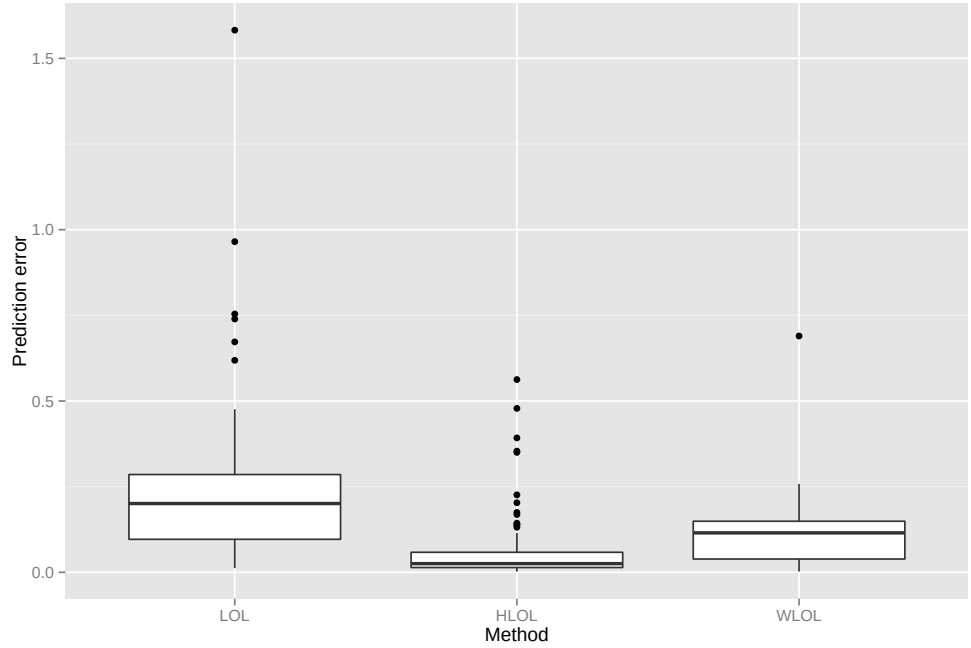

FIGURE 2.4. Normalized prediction error of the HLOL, LOL and WLOL procedures with autocorrelated noise, with covariance Γ^a, $a = 0.99$, deduced from 200 trials. SNR $= 5$

until the sparsity ratio becomes ≥ 0.2. After this point weighted adaptive Lasso is better. In the case of exponential noise the Lasso is almost always better except for extremely low sparsity ratio ($\leq 5\%$). This deficiency is partly explained by the fact that if LOL under performs the exponential noise will seriously increase the impact when estimating Γ.

As a conclusion OOSG should be considered for use when the user seeks:

— simplicity of implementation. Indeed OOSG is very straightforward to implement, relying on the most basic linear algebra libraries.

— speed. OOSG is very fast, on our benchmarks it can be three times faster than a lasso (or an adaptive lasso if we count in the time needed to compute the initial lasso estimator) as implemented for example in glmnet.

— when the covariance matrix of the noise is known, but very expensive, or very unstable to inverse, for example when there is long range correlation. In this case the transformed model behaves very badly.

— when it is reasonable to suppose that the signal is very sparse. Indeed in this case OOSG performs very well, and the price we pay in terms of accuracy compared to a more sophisticated method like the weighted adaptive lasso can be neglected compared to the important speed gain.

— when the covariance structure is complicated, and not just diagonal. In that case, unless we transform our initial model (and pay the price of the covariance inversion, in particular when this covariance is estimated), we do not know straightforward modifications of classical penalized estimators which would suit this situation.

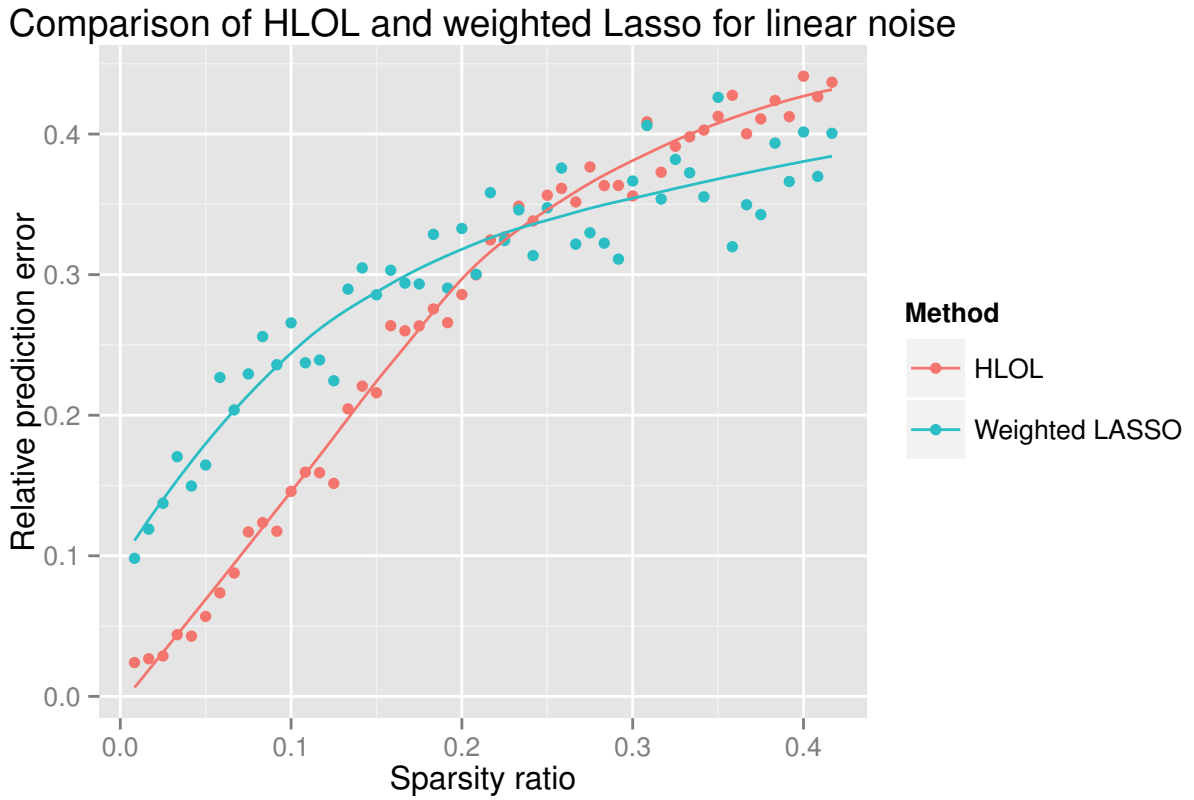

FIGURE 2.5. Comparison of the normalized predictive error of HLOL and weighted adaptive Lasso as a function of the sparsity ratio with a variance function $\sigma(\psi^i, \alpha) = | < \psi^i, \alpha > |$ from 200 trials. SNR = 5

6 Proofs

6.1 Preliminaries

Let S be a subset of indices, $S \subset \{1, \dots, p\}$. Define the least squares estimator restricted to S, which we denote $\hat{\alpha}(S)$, as a solution of:

$$\Psi_S \hat{\alpha}(S) = P_{V_S}[Z].$$

As soon as $|S| \leq \mathbf{N}_{\max}$, since the matrix ${}^t\Psi_S\Psi_S$ is invertible, there is a unique solution to the restricted least squares problem:

$$\hat{\alpha}(S) = ({}^t\Psi_S\Psi_S)^{-1}\,{}^t\Psi_S Z.$$

Given such a subset S, $|S| \leq \mathbf{N}_{\max}$, we denote by $\bar{\alpha}(S)$ the mean of $\hat{\alpha}(S)$:

$$\bar{\alpha}(S) = ({}^t\Psi_S\Psi_S)^{-1}\,{}^t\Psi_S\Psi\alpha = \alpha_S + ({}^t\Psi_S\Psi_S)^{-1}\,{}^t\Psi_S\Psi_{S^c}\alpha_{S^c}.$$

The estimator $\hat{\alpha}(S)$ can now be written as a sum of deterministic (at least conditionally on S) and a random term as:

$$\hat{\alpha}(S) = \bar{\alpha}(S) + ({}^t\Psi_S\Psi_S)^{-1}\,{}^t\Psi_S\eta.$$

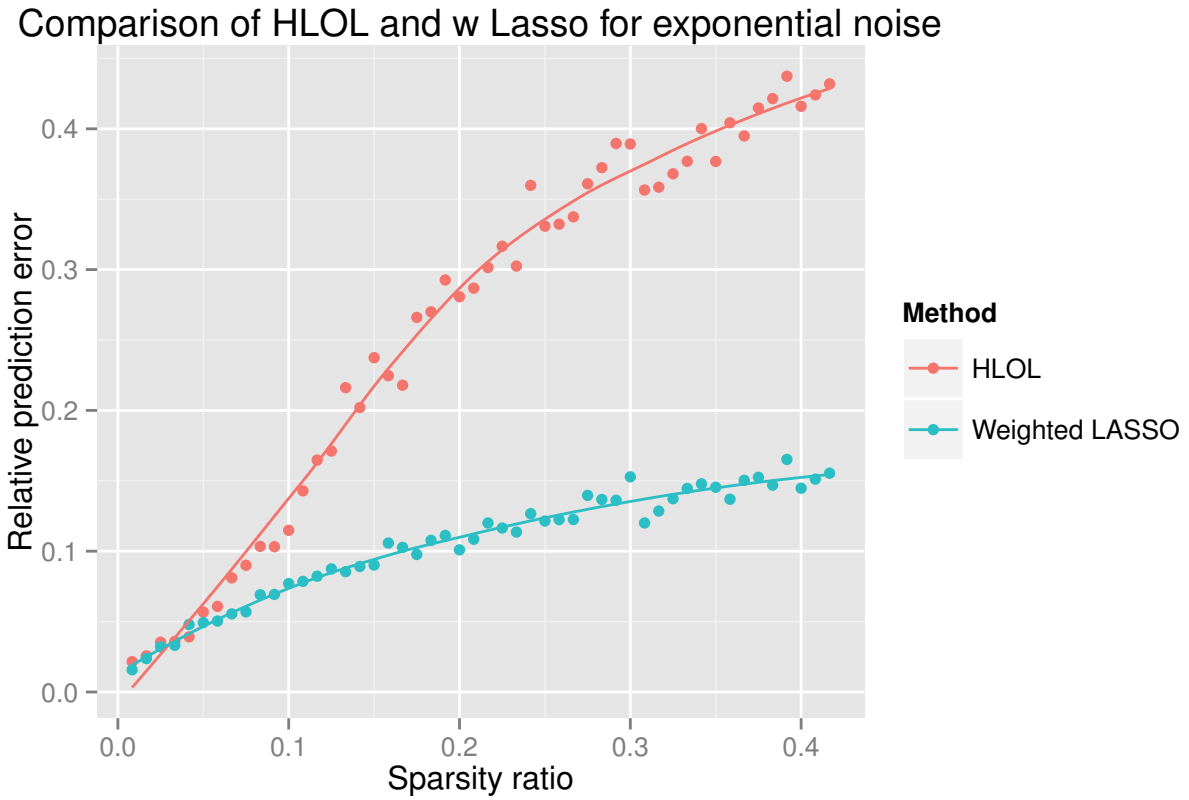

FIGURE 2.6. Comparison of the normalized predictive error of HLOL and weighted adaptive Lasso as a function of the sparsity ratio with a variance function $\sigma(\psi^i, \alpha) = \mathrm{e}^{|<\psi^i,\alpha>|}$ from 200 trials. SNR = 5

The next two propositions develop the necessary linear algebra to control the estimation error. First of all we bound the squared norm of the difference between the estimator $\hat{\alpha}(S)$ and the signal restricted to the subset of indices S, α_S. It is a measure of how well $\hat{\alpha}(S)$ approximates α_S.

Proposition 2.6. *Let S be a subset of indices such that $|S| \leq \mathbf{N}_{\max}$. Then:*

$$\|\alpha_S - \hat{\alpha}(S)\|_2^2 \leq 2\Big\{\Big(\frac{1}{1-\nu}\Big)^2 |S|\tau_n^2 \|\alpha_{S^c}\|_1^2 + \frac{1}{1-\nu}\frac{1}{n}\|P_{V_S}[\eta]\|_2^2\Big\}. \tag{2.8}$$

For a proof we refer to section 7.3. This proposition makes explicit the necessary bias coming from the restriction to a subset S in the term $|S|\tau_n^2\|\alpha_{S^c}\|_1^2$.

Starting from a restricted least squares estimator, $\hat{\alpha}(S)$, our procedure applies a final threshold, which is equivalent, given some subset $I \subset S$, to consider the estimator $\hat{\alpha}(S)_I$. The next proposition provides us with an algebraic bound on how well $\hat{\alpha}(S)_I$ approximates the signal α_I.

Proposition 2.7. *Let S be a subset of indices of cardinality bounded by $\mathbf{N}_{\max}$. Let I be a subset of S. Then there exists a constant C such that:*

$$\|\alpha_I - \hat{\alpha}(S)_I\|_2^2 \leq C\Big\{\Big[1 + (|S|\tau_n)^2\Big]|I|\tau_n^2\|\alpha_{I^c}\|_1^2 + \frac{1}{n}\|P_{V_I}[\eta]\|_2^2 + |I|\,|S|\tau_n^2\frac{1}{n}\|P_{V_S}[\eta]\|_2^2\Big\}.$$

We may take $C = \frac{4}{(1-\nu)^4}$.

For a proof we refer to section 7.4.

In both the preceding propositions, our bounds naturally incorporate the noise η. More precisely, in both, the squared norm of the orthogonal projection of the noise on some subspace V_S appears.

In the white noise case, where $\Gamma = \sigma^2 I_n$, only the dimension of the subspace V_S is relevant. In a general colored noise linear model, not only the dimension, but more generally the position of the subspace V_S can impact drastically the size of $\|P_{V_S}[\eta]\|_2^2$. Indeed consider for example the one dimensional subspaces $V_{\{i\}}$, where the associated covariates ψ_i would be the eigenvectors of Γ. Bounding those terms requires then to take into account an interaction between the design, the covariance of the noise and the particular subspaces V_S we consider in the procedure. Our main tool to handle the interaction between the design and the covariance matrix Γ will be the Γ-coherence of the design:

$$\tau_n(\Gamma) = \max_{l \neq l'} \frac{| < \psi_l, \Gamma \psi_{l'} > |}{\|\psi_l\|_\Gamma \|\psi_{l'}\|_\Gamma}.$$

Then we will only consider the random subsets $\mathbf{L}$ that can be selected by the procedure. As a consequence we will restrict ourselves to the random subsets of indices L belonging to $L^\lambda_{\Sigma_*,N}$, $\lambda > 0$ and $1 \leq N \leq \mathbf{N}_{\max}$. A subset of indices L belonging to $L^\lambda_{\Sigma_*,N}$ verifies the assumptions:

$$\forall\, l \in L, \quad |\check{\alpha}_l/\sigma_l| \geq \lambda, \tag{2.9}$$

where $\check{\alpha} = \frac{1}{n}{}^t\Psi Z$,

$$\sigma^2(L) \leq \Sigma_*, \tag{2.10}$$

and

$$|L| \leq N. \tag{2.11}$$

On those particular subsets we are able to bound the expectation of the orthogonal projection of the noise as summarized in the next proposition. Remark that the multiset $L^\lambda_{\Sigma_*,N}$ is attached to the model eq. (2.1).

Proposition 2.8. *Consider a linear model verifying the normalization eq.* (2.2).

1. *Let I be a deterministic subset of indices such that $1 \leq |I| \leq \mathbf{N}_{\max}$. Then:*

$$\mathbb{E}\Big[\|P_{V_I}[\eta]\|_2^2\Big] \leq \frac{1}{1-\nu}\sigma^2_{tot}(I)$$

and

$$\mathbb{E}\Big[\|P_{V_I}[\eta]\|_2^4\Big] \leq 1152\Big(\frac{1}{1-\nu}\Big)^2\sigma^4_{tot}(I),$$

where $\sigma^2_{tot}(I) = \frac{1}{n}\sum_{l\in I}\|\psi_l\|^2_\Gamma$.

2. *Consider a random subset of indices L of $L^{\lambda}_{\Sigma_*,N}$. Suppose that $\Sigma_* \leq p^{\theta}$, for some $\theta > 0$. Then for $p \geq 2$,*

$$\mathbb{E}[\|P_{V_L}[\eta]\|_2^2] \leq \frac{64}{1-\nu}(\sigma^2_{\max}(S) + \tau_n(\Gamma)\Sigma_*)N \log p,$$

for all $\lambda^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$, for some constant C depending on θ, where $S = S_{\frac{\lambda}{2}\sigma}(\alpha)$, and $\sigma^2_{\max}(S) = \max_{l\in S}\frac{\|\psi_l\|_\Gamma^2}{n} \vee 1$.

For a proof we refer to section 7.5.

6.2 The prediction error

The estimator produced by OOSG, $\hat{\alpha}(\mathbf{L}, \lambda_2)$, is necessarily supported on the initially selected set $\mathbf{L}$: $S_0(\hat{\alpha}(\mathbf{L}, \lambda_2)) \subset \mathbf{L}$. We may then bound the prediction error as a sum of two terms:

$$\mathbb{E}\Big[\frac{1}{n}\|\Psi(\alpha - \hat{\alpha})\|_2^2\Big] \leq 2\Big(\underbrace{\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}}(\alpha_{\mathbf{L}} - \hat{\alpha})\|_2^2\Big]}_{\text{Estimation error}} + \underbrace{\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}^c}\alpha_{\mathbf{L}^c}\|_2^2\Big]}_{\text{Selection error}}\Big).$$

The selection error results from the dimensionality reduction performed by the initial selection step. It measures the cost of not estimating the signal on the set of unselected covariates $\mathbf{L}^c$.

The estimation error measures how well the final estimator performs, for the prediction error, when estimating the restriction $\alpha_{\mathbf{L}}$ of the true signal α.

We will prove theorem 2.4 in two parts, bounding separately those two terms, which will give us theorem 2.4 as a result.

6.3 Selection error

Proposition 2.9. *Let $\lambda_1 > 0$ and consider the deterministic subset of indices $S = S_{\lambda_1\sigma}(\alpha)$. Suppose that λ_1 is large enough to verify $|S_{\frac{\lambda_1}{2}\sigma}(\alpha)| \leq \mathbf{N}_{\max}$. Then as soon as we choose $\Sigma \geq \sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha))$ and $\mathbf{N} \geq |S_{\frac{\lambda_1}{2}\sigma}(\alpha)|$, we have, if $p \geq 2$:*

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\boldsymbol{L}^c}\alpha_{\boldsymbol{L}^c}\|_2^2\Big] \leq & 4(1+\nu)\Big(\|\alpha_{S^c}\|_2^2 + \Big(\frac{\tau_n}{\nu} \wedge 1\Big)\|\alpha_{S^c}\|_1^2\Big) \\ & + 5|S|\tau_n^2\|\alpha\|_1^2 \\ & + 37\Big(\tau_n^2\|\alpha\|_1^2 + 5\frac{\log p}{n} + \lambda_1^2\Big)\sigma^2(S).\end{aligned}$$

Proof. The idea of the proof is to differentiate, among the unselected indices, those belonging to some set S, and those exterior to S, where S is the subset of indices where the

signal is mostly supported. Let $S = S_{\lambda_1\sigma}(\alpha) = \{l; |\alpha_l| \geq \sigma_l\lambda_1\}$. Then S is a deterministic subset, depending only on the signal of interest α, where the signal is larger than λ_1 times the noise level in the direction ψ_l.

We distinguish among the unselected indices $\mathbf{L}^c$ those belonging to S and those exterior to S:

$$\mathbf{L}^c = \underbrace{\left(\mathbf{L}^c \cap S^c_{\lambda_1\sigma}\right)}_{\mathbf{L}^c_e} \cup \underbrace{\left(\mathbf{L}^c \cap S_{\lambda_1\sigma}\right)}_{\mathbf{L}^c_i}.$$

So that $\mathbf{L}^c_e$ is the subset of indices the selection procedure was right not to keep, while $\mathbf{L}^c_i$ is where the selection procedure failed. We then decompose the selection error according to that splitting of the unselected indices:

$$\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}^c}\alpha_{\mathbf{L}^c}\|_2^2\Big] \leq 2\Big(\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}^c_e}\alpha_{\mathbf{L}^c_e}\|_2^2\Big] + \mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}^c_i}\alpha_{\mathbf{L}^c_i}\|_2^2\Big]\Big).$$

We handle both of the terms on the right separately.

To handle $\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}^c_e}\alpha_{\mathbf{L}^c_e}\|_2^2\Big]$ we cannot use directly lemma 2.1 since $\mathbf{L}^c_e$ can be of cardinality much larger than $\mathbf{N}_{\max}$. But the following proposition from [76] bounds how much a design matrix Ψ, satisfying a restricted isometry property, inflates non-sparse vectors.

Proposition 2.10. *Suppose that a $n \times p$ matrix Φ verifies, for some constant $\delta_r > 0$,*

$$\|\Phi x\|_2 \leq \sqrt{1+\delta_r}\|x\|_2,$$

for all r-sparse vector x, i.e x has at most r non zero coefficients. Then

$$\forall x \in \mathbb{R}^p, \quad \|\Phi x\|_2 \leq \sqrt{1+\delta_r}\Big[\|x\|_2 + \frac{1}{\sqrt{r}}\|x\|_1\Big].$$

According to lemma 2.1 we know that $\frac{1}{\sqrt{n}}\Psi$ is such that:

$$\Big\|\frac{1}{\sqrt{n}}\Psi x\Big\|_2 \leq \sqrt{1+\nu}\|x\|_2,$$

for all $x \in \mathbb{R}^p$ which are $\mathbf{N}_{\max}$-sparse. So that according to proposition 2.10:

$$\begin{aligned}\Big\|\frac{1}{\sqrt{n}}\Psi_{\mathbf{L}^c_e}\alpha_{\mathbf{L}^c_e}\Big\|_2 &\leq \sqrt{1+\nu}\Big[\|\alpha_{\mathbf{L}^c_e}\|_2 + \frac{1}{\sqrt{\mathbf{N}_{\max}}}\|\alpha_{\mathbf{L}^c_e}\|_1\Big],\\ &\leq \sqrt{1+\nu}\Big[\|\alpha_{S^c}\|_2 + \frac{1}{\sqrt{\mathbf{N}_{\max}}}\|\alpha_{S^c}\|_1\Big].\end{aligned}$$

Furthermore according to lemma 2.2, $\mathbf{N}_{\max} \geq \nu/\tau_n \vee 1$, which implies that:

$$\mathbb{E}\Big[\frac{1}{n}\Big\|\Psi_{\mathbf{L}^c_e}\alpha_{\mathbf{L}^c_e}\Big\|_2^2\Big] \leq 2(1+\nu)\Big[\|\alpha_{S^c}\|_2^2 + \Big(\frac{\tau_n}{\nu} \wedge 1\Big)\|\alpha_{S^c}\|_1^2\Big]. \tag{2.12}$$

We can notice that this bound would be greatly improved by any tighter lower bound on $\mathbf{N}_{\max}$. Furthermore we can notice that this bound does not depend on the particular

set S relatively to which we split $\mathbf{L}^c$.

It remains to bound $\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\mathbf{L}_i^c}\alpha_{\mathbf{L}_i^c}\|_2^2\Big]$. Suppose that we take λ_1 large enough to verify $|S| \leq \mathbf{N}_{\max}$. Then since $\mathbf{L}_i^c \subset S$, we can apply lemma 2.1:

$$\frac{1}{n}\|\Psi_{\mathbf{L}_i^c}\alpha_{\mathbf{L}_i^c}\|_2^2 \leq (1+\nu)\|\alpha_{\mathbf{L}_i^c}\|_2^2.$$

We now have to bound $\|\alpha_{\mathbf{L}_i^c}\|_2$. Assuming that the selection step is not to wrong is equivalent to assuming that the initial estimator $\check{\alpha}$ is not too far from α, at least on S. It it then natural to proceed through the triangular inequality:

$$\begin{aligned}
\|\alpha_{\mathbf{L}_i^c}\|_2^2 &= \sum_{l\in\mathbf{L}_i^c}(\alpha_l - \check{\alpha}_l + \check{\alpha}_l)^2, \\
&\leq 2\Big[\sum_{l\in\mathbf{L}_i^c}(\alpha_l-\check{\alpha}_l)^2 + \sum_{l\in\mathbf{L}_i^c}\check{\alpha}_l^2\Big], \\
&\leq 2\Big[\sum_{l\in S}(\alpha_l-\check{\alpha}_l)^2 + \sum_{l\in\mathbf{L}_i^c}\check{\alpha}_l^2\Big].
\end{aligned}$$

Furthermore since we know the distribution of $\check{\alpha}_l$, from lemma 2.3, we have that:

$$\mathbb{E}\Big[\sum_{l\in S}(\alpha_l-\check{\alpha}_l)^2\Big] = \sum_{l\in S}R_l^2 + \sum_{l\in S}\frac{\|\psi_l\|_\Gamma^2}{n^2}.$$

It remains to handle the sum $\sum_{l\in\mathbf{L}_i^c}\check{\alpha}_l^2$ in expectation. To do, following a classical trick, we split $\mathbf{L}_i^c$ into $\mathbf{L}_i^c = K_1 \cup K_2$, where:

$$K_1 = \mathbf{L}_i^c \cap \{l; |\check{\alpha}_l| \leq 2|\alpha_l - \check{\alpha}_l|\}$$

and

$$K_2 = \mathbf{L}_i^c \cap \{l; |\check{\alpha}_l| > 2|\alpha_l - \check{\alpha}_l|\},$$

and split the whole sum $\sum_{l\in\mathbf{L}_i^c}\check{\alpha}_l^2$ accordingly:

$$\sum_{l\in\mathbf{L}_i^c}\check{\alpha}_l^2 = \sum_{l\in K_1}\check{\alpha}_l^2 + \sum_{l\in K_2}\check{\alpha}_l^2.$$

By construction the sum on K_1 is handled in a straightforward manner:

$$\mathbb{E}\Big[\sum_{l\in K_1}\check{\alpha}_l^2\Big] \leq 4\Big\{\sum_{l\in S}R_l^2 + \sum_{l\in S}\frac{\|\psi_l\|_\Gamma^2}{n^2}\Big\}.$$

The sum on K_2 is more subtle to handle. Reconsider the way the set $\mathbf{L}$ is selected. Recall that $S_{\lambda_1\sigma}(\check{\alpha}) = \{l; |\check{\alpha}_l/\sigma_l| \geq \lambda_1\}$, which is a random but observable set (it can be computed from the data). We pick covariates ψ_l, for $l \in S_{\lambda_1\sigma}(\check{\alpha})$, in the order of the normalized correlations $|\check{\alpha}_l/\sigma_l|$ from the largest to the smallest, until adding one more would make the sum $\sum_{l\in\mathbf{L}}\sigma_l^2$ larger than some a priori bound Σ, or that we have exhausted the set $S_{\lambda_1\sigma}(\check{\alpha})$.

Consider the set $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$, which a slight enlargement of S, $S \subset S_{\frac{\lambda_1}{2}\sigma}(\alpha)$. Then we have to consider the following cases:

— we have selected all the indices belonging to $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$, $S_{\frac{\lambda_1}{2}\sigma}(\alpha) \subset \mathbf{L}$, but in that case we have selected all the indices of S, $\mathbf{L}_i^c = \emptyset$, and there is nothing to prove,

— $\mathbf{L}$ is a proper subset of $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$ and we have exhausted $S_{\lambda_1\sigma}(\check{\alpha})$, so that for all $l \in K_2$, $|\check{\alpha}_l| \leq \lambda_1\sigma_l$. Then:

$$\sum_{l\in K_2} \check{\alpha}_l^2 \leq \lambda_1^2 \sum_{l\in K_2} \sigma_l^2 \leq \lambda_1^2 \sum_{l\in S} \sigma_l^2.$$

— $\mathbf{L}$ is a proper subset of $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$ and $S_{\lambda_1\sigma}(\check{\alpha})$ has not been exhausted. Then, since we suppose λ_1 to verify the inequality $\sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) \leq \Sigma$, by maximality of $\mathbf{L}$ there must exist an index $l' \in S^c_{\frac{\lambda_1}{2}\sigma}(\alpha)$ such that:

$$\forall l \in K_2, \quad |\check{\alpha}_{l'}/\sigma_{l'}| \geq |\check{\alpha}_l/\sigma_l|,$$

which cannot be added to $\mathbf{L}$, since it would make $\sigma^2(\mathbf{L})$ get larger than Σ.

— $\mathbf{L}$ is not a subset of $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$. Then, if $\mathbf{L}_i^c$ is not empty, there exists indices $l' \in \mathbf{L}$ and $l' \notin S_{\frac{\lambda_1}{2}\sigma}(\alpha)$.

In the last two cases, if K_2 is not empty, then necessarily for each $l \in K_2$, there exists an obstruction index $l^*(l)$ such that:

1. its normalized correlation with the observed signal Z is larger than l's,

$$|\check{\alpha}_{l^*(l)}/\sigma_{l^*(l)}| \geq |\check{\alpha}_l/\sigma_l|,$$

2. and it is exterior to $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$, $|\alpha_{l^*(l)}/\sigma_{l^*(l)}| < \frac{\lambda_1}{2}$.

Furthermore by construction of K_2, for all $l \in K_2$,

$$|\alpha_l| - |\check{\alpha}_l| \leq |\alpha_l - \check{\alpha}_l| < \frac{1}{2}|\check{\alpha}_l|,$$

so that $|\alpha_l| < \frac{3}{2}|\check{\alpha}_l|$. The signal on K_2 is dominated in size by a constant times the initial estimation $\check{\alpha}$ on K_2.

And by construction of the obstruction index, since $l^*(l) \notin S_{\frac{\lambda_1}{2}\sigma}(\alpha)$, for all $l \in K_2$:

$$|\alpha_{l^*(l)}/\sigma_{l^*(l)}| < \lambda_1/2 < \frac{1}{2}|\alpha_l/\sigma_l|.$$

Finally, adding those considerations:

$$\begin{aligned}
|\check{\alpha}_{l^*(l)}/\sigma_{l^*(l)} - \alpha_{l^*(l)}/\sigma_{l^*(l)}| &\geq |\check{\alpha}_{l^*(l)}/\sigma_{l^*(l)}| - |\alpha_{l^*(l)}/\sigma_{l^*(l)}|, \\
&\geq |\check{\alpha}_l/\sigma_l| - \frac{1}{2}|\alpha_l/\sigma_l|, \\
&\geq |\check{\alpha}_l/\sigma_l| - \frac{3}{4}|\check{\alpha}_l/\sigma_l|, \\
&\geq \frac{1}{4}|\check{\alpha}_l/\sigma_l|.
\end{aligned}$$

To summarize, we proved that if $S_{\lambda_1\sigma}(\check{\alpha})$ has not been exhausted by the first selection step and $K_2 \neq \emptyset$, for each $l \in K_2$ there exists an index $1 \leq l^*(l) \leq p$, such that:

$$|\check{\alpha}_l| \leq 4\frac{\sigma_l}{\sigma_{l^*(l)}}|\check{\alpha}_{l^*(l)} - \alpha_{l^*(l)}|. \tag{2.13}$$

Using for all $l \in K_2$ this construction of an obstruction index $l^*(l)$, we get:

$$\sum_{l\in K_2} \check{\alpha}_l^2 \leq 16\Big[\sum_{l\in K_2} (\check{\alpha}_{l^*(l)} - \alpha_{l^*(l)})^2 \frac{\sigma_l^2}{\sigma_{l^*(l)}^2}\Big].$$

We recall that for all indices $1 \leq l \leq p$, $\eta_l = \frac{1}{n} < \psi_l, \eta >= \frac{\|\psi_l\|_\Gamma}{n}\tilde{\eta}_l$, where $\tilde{\eta}_l \sim \mathcal{N}(0,1)$. Then using lemma 2.3 we get for all $1 \leq k \leq p$:

$$(\check{\alpha}_k - \alpha_k)^2 \leq 2(R_k^2 + \frac{\|\psi_k\|_\Gamma^2}{n^2}\tilde{\eta}_k^2) \leq 2(R_k^2 + \frac{\sigma_k^2}{n}\tilde{\eta}_k^2).$$

So that for all $l \in K_2$ we have:

$$\begin{aligned}
\mathbb{E}\Big[(\check{\alpha}_{l^*(l)} - \alpha_{l^*(l)})^2 \frac{\sigma_l^2}{\sigma_{l^*(l)}^2}\Big] &= \mathbb{E}\Big[\sum_{k=1}^{p} (\check{\alpha}_k - \alpha_k)^2 \frac{\sigma_l^2}{\sigma_k^2}\, \mathbb{1}\{l^*(l) = k\}\Big], \\
&\leq 2\sigma_l^2 \mathbb{E}\Big[\sum_{k=1}^{p} \frac{R_k^2}{\sigma_k^2}\, \mathbb{1}\{l^*(l) = k\} + \frac{1}{n}\sum_{k=1}^{p} \tilde{\eta}_k^2\, \mathbb{1}\{l^*(l) = k\}\Big], \\
&\leq 2\sigma_l^2 \Big(\tau_n^2 \|\alpha\|_1^2 + \frac{1}{n}\mathbb{E}[\max_{1\leq k\leq p} \tilde{\eta}_k^2]\Big).
\end{aligned}$$

The random variables $\tilde{\eta}_k^2$ have a chi-squared distribution with one degree of freedom, and it is well known [80], [10], that for $p \geq 2$:

$$\mathbb{E}[\max_{1\leq k\leq p} \tilde{\eta}_k^2] \leq 1 + 2\sqrt{\log(p)} + 2\log(p) \leq 5\log(p).$$

This finally allows us to bound the initial sum in expectation in the case where K_2 is not empty and $S_{\lambda_1\sigma}(\check{\alpha})$ has not been exhausted:

$$\mathbb{E}\Big[\sum_{l\in K_2} \check{\alpha}_l^2\Big] \leq 32\Big(\tau_n^2\|\alpha\|_1^2 + 5\frac{log(p)}{n}\Big)\sum_{l\in S} \sigma_l^2, \tag{2.14}$$

which concludes the proof. □

6.4 Estimation error

Proposition 2.11. *Let $K_1, K_2, \theta > 0$ be constants. If we choose the parameters of the procedure such that:*

1. $\tau_n N \leq K_1$ *and* $\tau_n(\Gamma)\Sigma \leq K_2$,
2. $\Sigma \leq p^\theta$, $\mathbf{N} \leq \mathbf{N}_{\max}$
3. $\lambda_1^2 \geq C_2\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$, *and*
4. $\lambda_2 \leq \lambda_1$

then:

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{n}\|\Psi_{\boldsymbol{L}}(\alpha_{\boldsymbol{L}} - \hat{\alpha}(\boldsymbol{L}, \lambda_2))\|_2^2\Big] \leq C_1\Big\{&\|\alpha_{S^c_{2\lambda_2\sigma}(\alpha)}\|_2^2 + \\
&+ \Big(\tau_n^2\|\alpha\|_1^2 + \sigma_{\max}^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big)\sigma^2\Big(S_{\frac{\lambda_2}{2}\sigma}(\alpha)\Big)\Big\}.
\end{aligned}$$

Proof. To bound the estimation error, which is the expectation of $I := \frac{1}{n}\|\Psi_{\mathbf{L}}(\alpha_{\mathbf{L}} - \hat{\alpha}(\mathbf{L}, \lambda_2))\|_2^2$, we start by splitting the error according to the thresholding procedure:

$$\begin{aligned}
I \leq\ & 8\Big(\frac{1}{n}\|\sum_{l\in\mathbf{L}}(\alpha_l - \hat{\alpha}(\mathbf{L})_l)\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| \geq \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| \geq \lambda_2/2\}\|_2^2 \\
& + \frac{1}{n}\|\sum_{l\in\mathbf{L}}(\alpha_l - \hat{\alpha}(\mathbf{L})_l)\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| \geq \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| < \lambda_2/2\}\|_2^2 \\
& + \frac{1}{n}\|\sum_{l\in\mathbf{L}}\alpha_l\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| > 2\lambda_2\}\|_2^2 \\
& + \frac{1}{n}\|\sum_{l\in\mathbf{L}}\alpha_l\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| \leq 2\lambda_2\}\|_2^2\Big), \\
\leq\ & 8\Big(BB + BS + SB + SS\Big),
\end{aligned}$$

where B denotes "Big" and S "Small".

Study of SS:

We first study SS which accounts for those indices selected among $\mathbf{L}$, but discarded by the second threshold, while the signal was indeed small on them:

$$\begin{aligned}
\text{SS} &:= \frac{1}{n}\Big\|\sum_{l\in\mathbf{L}}\alpha_l\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\}\mathbb{1}\{|\alpha_l/\sigma_l| \leq 2\lambda_2\}\Big\|_2^2, \\
&= \frac{1}{n}\|\Psi_{S_1}\alpha_{S_1}\|_2^2,
\end{aligned}$$

where $S_1 = \mathbf{L} \cap \{l; |\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\} \cap \{l; |\alpha_l/\sigma_l| \leq 2\lambda_2\}$.

If $|\mathbf{L}| \leq \mathbf{N}_{\max}$, then since $S_1 \subset \mathbf{L}$, $|S_1| \leq \mathbf{N}_{\max}$, so that we can apply lemma 2.1:

$$\begin{aligned}
SS &\leq (1+\nu)\Big[\sum_{l\in\mathbf{L}}\alpha_l^2\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| \leq 2\lambda_2\}\Big], \\
&\leq (1+\nu)\Big[\sum_{l\in\mathbf{L}}\alpha_l^2\, \mathbb{1}\{|\alpha_l/\sigma_l| \leq 2\lambda_2\}\Big], \\
&\leq (1+\nu)\Big[\sum_{l=1}^{p}\alpha_l^2\, \mathbb{1}\{|\alpha_l/\sigma_l| \leq 2\lambda_2\}\Big].
\end{aligned}$$

Finally:

$$SS \leq (1+\nu)\|\alpha_{S^c_{2\lambda_2\sigma}(\alpha)}\|_2^2, \tag{2.15}$$

where we recall that $S^c_{2\lambda_2\sigma}(\alpha) = \{l; |\alpha_l| \leq 2\sigma_l\lambda_2\}$.

Study of BB and SB:

We recall that BB is the term accounting for those columns selected among the leaders, not thresholded and where the signal is "big":

$$\text{BB} := \frac{1}{n}\Big\|\sum_{l\in\mathbf{L}}(\alpha_l - \hat{\alpha}(\mathbf{L})_l)\psi_l\, \mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| \geq \lambda_2\}\, \mathbb{1}\{|\alpha_l/\sigma_l| \geq \lambda_2/2\}\Big\|_2^2.$$

On the other hand SB is the term accounting for the columns discarded by the second threshold despite the signal being "big":

$$\mathrm{SB} := \frac{1}{n}\Big\| \sum_{l\in\mathbf{L}} \alpha_l\psi_l \,\mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| < \lambda_2\}\,\mathbb{1}\{|\alpha_l/\sigma_l| > 2\lambda_2\}\Big\|_2^2.$$

We study those two terms together since, as a first step we show that SB and BB are essentially bounded by the same quantity. Indeed:

$$\mathbb{1}\{|\hat{\alpha}_l/\sigma_l| < \lambda_2\}\,\mathbb{1}\{|\alpha_l/\sigma_l| > 2\lambda_2\} \le \mathbb{1}\{|\hat{\alpha}_l/\sigma_l| < \lambda_2 < |\hat{\alpha}_l/\sigma_l - \alpha_l/\sigma_l|\}\,\mathbb{1}\{|\alpha_l/\sigma_l| > \lambda_2/2\}.$$

As a consequence, using lemma 2.1:

$$\begin{aligned}
\mathrm{SB} &\le (1+\nu)\Big[\sum_{l\in\mathbf{L}} \alpha_l^2\,\mathbb{1}\{|\hat{\alpha}_l/\sigma_l| < \lambda_2 < |\hat{\alpha}_l/\sigma_l - \alpha_l/\sigma_l|\}1\{|\alpha_l/\sigma_l| > \lambda_2/2\}\Big],\\
&\le 2(1+\nu)\Big(\sum_{l\in\mathbf{L}} (\alpha_l - \hat{\alpha}_l)^2\,\mathbb{1}\{|\alpha_l/\sigma_l| > \lambda_2/2\}\\
&\quad + \sum_{l\in\mathbf{L}} \hat{\alpha}_l^2\,\mathbb{1}\{|\hat{\alpha}_l/\sigma_l| < \lambda_2 < |\hat{\alpha}_l/\sigma_l - \alpha_l/\sigma_l|\}1\{|\alpha_l/\sigma_l| > \lambda_2/2\}\Big),\\
&\le 4(1+\nu)\Big[\sum_{l\in\mathbf{L}} (\alpha_l - \hat{\alpha}_l)^2\,\mathbb{1}\{|\alpha_l/\sigma_l| > \lambda_2/2\}\Big].
\end{aligned}$$

Even more directly we get that:

$$\mathrm{BB} \le (1+\nu)\Big[\sum_{l\in\mathbf{L}} (\alpha_l - \hat{\alpha}_l)^2\,\mathbb{1}\{|\alpha_l/\sigma_l| > \lambda_2/2\}\Big].$$

In the end:

$$\mathrm{SB} + \mathrm{BB} \le 5(1+\nu)\sum_{l\in\mathbf{L}} (\alpha_l - \hat{\alpha}(\mathbf{L})_l)^2\,\mathbb{1}\{|\alpha_l/\sigma_l| > \lambda_2/2\}.$$

Notice that $S_{\frac{\lambda_2}{2}\sigma}(\alpha)$ is a deterministic subset of indices, and

$$S_2 = \mathbf{L} \cap S_{\frac{\lambda_2}{2}\sigma}(\alpha)$$

is a random one. We can then rewrite:

$$\mathrm{SB} + \mathrm{BB} \le 5(1+\nu)\|\alpha_{S_2} - \hat{\alpha}(\mathbf{L})_{S_2}\|_2^2.$$

Then using proposition 2.7 we have:

$$\begin{aligned}
\|\alpha_{S_2} - \hat{\alpha}_{S_2}\|_2^2 &\le C\Big\{[1 + (|\mathbf{L}|\tau_n)^2]|S_2|\tau_n^2\|\alpha_{S_2^c}\|_1^2 + \frac{1}{n}\|P_{V_{S_2}}[\eta]\|_2^2 + |S_2||\mathbf{L}|\tau_n^2\frac{1}{n}\|P_{V_{\mathbf{L}}}[\eta]\|_2^2\Big\},\\
&\le C\Big\{[1 + (N\tau_n)^2]|S_{\frac{\lambda_2}{2}\sigma}(\alpha)|\tau_n^2\|\alpha\|_1^2 + \frac{1}{n}\|P_{V_{S_{\frac{\lambda_2}{2}\sigma}(\alpha)}}[\eta]\|_2^2\\
&\quad + |S_{\frac{\lambda_2}{2}\sigma}(\alpha)|N\tau_n^2\frac{1}{n}\|P_{V_{\mathbf{L}}}[\eta]\|_2^2\Big\}.
\end{aligned}$$

According to proposition 2.8 we have:

$$\mathbb{E}[\|P_{V_{S_{\frac{\lambda_2}{2}\sigma}(\alpha)}}[\eta]\|_2^2] \leq \frac{1}{1-\nu}\sigma^2\Big(S_{\frac{\lambda_2}{2}\sigma}(\alpha)\Big), \tag{2.16}$$

and

$$\mathbb{E}[\|P_{V_{\mathbf{L}}}[\eta]\|_2^2] \leq \frac{64}{1-\nu}(\sigma^2_{\max}(S_{\frac{\lambda_2}{2}\sigma}(\alpha)) + \tau_n(\Gamma)\Sigma)N\log p, \tag{2.17}$$

for $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$.

As a result:

$$\mathbb{E}[\|\alpha_{S_2} - \hat{\alpha}(\mathbf{L})_{S_2}\|_2^2] \leq \frac{64C}{1-\nu}\Big\{[1+(N\tau_n)^2]\tau_n^2\|\alpha\|_1^2 + \frac{1}{n} + \Big(\sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) + \tau_n(\Gamma)\Sigma\Big)(N\tau_n)^2\frac{\log p}{n}\Big\}\sigma^2\Big(S_{\frac{\lambda_2}{2}\sigma}(\alpha)\Big),$$

as soon as $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$.

Now since we suppose that $N\tau_n \leq K_1$ and $\tau_n(\Gamma)\Sigma \leq K_2$, there exists a constant D, depending on K_1 and K_2, such that:

$$\mathbb{E}[\mathrm{SB}+\mathrm{BB}] \leq D\Big\{\|\alpha\|_1^2\tau_n^2 + \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big\}\sigma^2\Big(S_{\frac{\lambda_2}{2}\sigma}(\alpha)\Big), \tag{2.18}$$

as soon as $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$, and $p \geq 2$.

Study of BS: Finally BS is the term accounting for those columns selected by the second threshold while the signal is "small":

$$\mathrm{BS} := \frac{1}{n}\Big\|\sum_{l\in\mathbf{L}}(\alpha_l - \hat{\alpha}(\mathbf{L})_l)\psi_l\,\mathbb{1}\{|\hat{\alpha}(\mathbf{L})_l/\sigma_l| \geq \lambda_2\}\,\mathbb{1}\{|\alpha_l/\sigma_l| < \lambda_2/2\}\Big\|_2^2.$$

Let $S_3 = \{l \in \mathbf{L}; |\hat{\alpha}_l/\sigma_l| \geq \lambda_2, |\alpha_l/\sigma_l| < \lambda_2/2\}$. Then using lemma 2.1 we have:

$$\mathrm{BS} \leq (1+\nu)\Big[\sum_{l\in S_3}(\alpha_l - \hat{\alpha}_l)^2\Big] = (1+\nu)\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2.$$

And:

$$\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2] = \underbrace{\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2\,\mathbb{1}\{\sigma^2(S_3) \leq \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))\}]}_{A} + \underbrace{\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2\,\mathbb{1}\{\sigma^2(S_3) > \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))\}]}_{B}.$$

Furthermore from proposition 2.7, and since $N\tau_n \leq K_1$ and $\tau_n(\Gamma)\Sigma \leq K_2$, there exists a constant D, depending on K_1 and K_2, such that:

$$\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2 \leq D\Big\{|S_3|\tau_n^2\|\alpha\|_1^2 + \frac{1}{n}\|P_{V_{S_3}}[\eta]\|_2^2 + |S_3|\tau_n\frac{1}{n}\|P_{V_{\mathbf{L}}}[\eta]\|_2^2\Big\},$$

And using proposition 2.8:

$$\mathbb{E}[\|P_{V_{\mathbf{L}}}[\eta]\|_2^2] \leq D' \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) N \log p,$$

as soon as $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$.
Finally:

$$\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2] \leq D'' \mathbb{E}\Big[|S_3|\tau_n^2\|\alpha\|_1^2 + \frac{1}{n}\|P_{V_{S_3}}[\eta]\|_2^2 + |S_3|\sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big],$$

for some constant D'', as soon as $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$.

Furthermore using proposition 2.8, on the event $\{\sigma^2(S_3) \leq \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))\}$, we have:

$$\mathbb{E}[\frac{1}{n}\|P_{V_{S_3}}[\eta]\|_2^2 \mathbb{1}\{\sigma^2(S_3) \leq \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))\}] \leq T\sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha)),$$

for some constant T. So that:

$$A \leq D_1\Big(\tau_n^2\|\alpha\|_1^2 + \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big)\sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha)), \tag{2.19}$$

for $\lambda_1^2 \geq C\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$, for some constant D_1.

It remains to handle B. From Cauchy-Schwarz:

$$\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^2 1\{\sigma^2(S_3) > \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))\}] \leq \sqrt{\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^4]}\sqrt{\mathbb{P}(\sigma^2(S_3) > \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))}.$$

Since $S_3 \subset \mathbf{L}$:

$$\sqrt{\mathbb{E}[\|\alpha_{S_3} - \hat{\alpha}(\mathbf{L})_{S_3}\|_2^4]} \leq 4D_1\Big(\tau_n^2\|\alpha\|_1^2 + \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big)\Sigma.$$

Now since $\lambda_2 < \lambda_1$, $\mathbb{P}(\sigma^2(S_3) > \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha))$ is inferior to the probability of having any index external to $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$ among $\mathbf{L}$. And the probability that a given index l belongs to $\mathbf{L}$ and not to $S_{\frac{\lambda_1}{2}\sigma}(\alpha)$ can be bounded in the following manner:

$$\begin{aligned}
\mathbb{P}(l \in \mathbf{L}, l \notin S_{\frac{\lambda_1}{2}\sigma}(\alpha)) &\leq \mathbb{P}(|\check{\alpha}_l/\sigma_l| \geq \lambda_1, |\alpha_l/\sigma_l| < \lambda_1/2) \\
&\leq \mathbb{P}(|\check{\alpha}_l - \alpha_l| \geq \sigma_l\lambda_1/2), \\
&\leq 2e^{-n\lambda_1^2/16},
\end{aligned}$$

for $\lambda_1 \geq 8\tau_n\|\alpha\|_1/\sigma_l$.

Then by union bound:

$$\mathbb{P}(\sigma^2(S_3) > \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha)) \leq 2pe^{-n\lambda_1^2/16},$$

as soon as $\lambda_1 \geq 8\tau_n\|\alpha\|_1$ (since for all l, $\sigma_l \geq 1$). Finally we get that:

$$B \leq 8D_1\Big(\tau_n^2\|\alpha\|_1^2 + \sigma^2_{\max}(S_{\frac{\lambda_1}{2}\sigma}(\alpha))\frac{\log p}{n}\Big)\Sigma p e^{-n\lambda_1^2/32}. \tag{2.20}$$

So that $B \leq A$ as soon as:

$$8\Sigma p e^{-n\lambda_1^2/32} \leq \sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha)),$$

which is realized as soon as $\lambda_1^2 \geq M\Big(\frac{\log p}{n} \vee (\tau_n\|\alpha\|_1)^2\Big)$, for some constant M depending on θ. □

Combining the two previous propositions, on the selection and the estimation error, we obtain the general theorem 2.4.

6.5 Proof of theorem 2.5

We prove theorem 2.5 as a corollary of theorem 2.4.

Proof. First of all we start by recalling the following facts: if $\alpha \in \mathcal{B}_{q,\sigma}(M)$, for $q \in (0,1]$, then

1. by Markov's inequality:

$$\forall \lambda > 0, \quad \sigma^2\Big(\{l; |\alpha_l/\sigma_l| \geq \lambda\}\Big) = \sum_{l=1}^{p} \sigma_l^2 1\{|\alpha_l/\sigma_l| \geq \lambda\} \leq M^q \lambda^{-q}, \tag{2.21}$$

2. for all $p \geq q$:

$$\forall \lambda \geq 0, \quad \sum_{l=1}^{p} \sigma_l^2 |\alpha_l/\sigma_l|^p 1\{|\alpha_l/\sigma_l| \leq \lambda\} \leq M^q \lambda^{p-q}. \tag{2.22}$$

— Suppose that $\alpha \in \mathcal{B}_{q,\sigma}(M)$, $q \in (0,1]$, then according to eq. (2.21):

$$\sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) \leq (2M)^q \lambda_1^{-q}.$$

Then as soon as $\lambda_1 \geq 2M\Big(1 \wedge \frac{\tau_n}{\nu}\Big)$, we have $\sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) \leq \frac{\nu}{\tau_n} \vee 1 \leq \mathbf{N}_{\max}$. We can then apply theorem 2.4. To do so we have to bound $\|\alpha_{S^c_{2\lambda_1\sigma}}\|_2$ and $\|\alpha_{S^c_{2\lambda_1\sigma}}\|_1$. Using eq. (2.22), and since $\alpha \in \mathcal{B}_{q,\sigma}(M)$ for some $q \in (0,1]$, we have that:

$$\begin{aligned}
\|\alpha_{S^c_{2\lambda_1\sigma}}\|_2^2 &= \sum_{l=1}^{p} \alpha_l^2 1\{|\alpha_l/\sigma_l| < 2\lambda_1\}, \\
&= \sum_{l=1}^{p} \sigma_l^2 (\alpha_l/\sigma_l)^2 1\{|\alpha_l/\sigma_l| < 2\lambda_1\}, \\
&\leq 4M^q \lambda_1^{2-q} \leq 4(1 \vee M)\lambda_1^{2-q}.
\end{aligned}$$

If $\alpha \in \mathcal{B}_{q,\sigma}(M)$ for some $q \in (0,1]$, we bound $\|\alpha_{S^c_{2\lambda_1\sigma}}\|_1$, using that $\sigma_l \geq 1$ for all l, and eq. (2.22):

$$\begin{aligned}
\|\alpha_{S^c_{2\lambda_1\sigma}}\|_1 &= \sum_{l=1}^{p} |\alpha_l| 1\{|\alpha_l/\sigma_l| < 2\lambda_1\}, \\
&\leq \sum_{l=1}^{p} \sigma_l^2 |\alpha_l/\sigma_l| 1\{|\alpha_l/\sigma_l| < 2\lambda_1\}, \\
&\leq 2M^q \lambda_1^{1-q}.
\end{aligned}$$

Since $\lambda_1 \geq 2M\Big(1 \wedge \frac{\tau_n}{\nu}\Big)$, we have:

$$\Big(\frac{\tau_n}{\nu} \wedge 1\Big)\|\alpha_{S^c_{2\lambda_1\sigma}}\|_1^2 \leq 8(1 \vee M)\lambda_1^{2-q}.$$

Finally by eq. (2.21):

$$\sigma^2(S_{\frac{\lambda_2}{2}\sigma}(\alpha)) \leq 2(1 \vee M)\lambda_2^{-q},$$

which gives the result.

— Suppose that $\alpha \in \mathcal{B}_{0,\sigma}(S, M)$, then:

$$\sigma^2(S_{\frac{\lambda_1}{2}\sigma}(\alpha)) \leq \sum_{l=1}^{p} \sigma_l^2 1\{\alpha_l \neq 0\} \leq S$$

and we suppose $S \leq \nu/\tau_n \vee 1$, so that we can apply theorem 2.4. And:

$$\|\alpha_{S^c_{2\lambda_1\sigma}}\|_2^2 \leq 4S\lambda_1^2,$$

$$\|\alpha_{S^c_{2\lambda_1\sigma}}\|_1 \leq 2S\lambda_1.$$

□

7 Appendix

7.1 Proof of lemma 2.1

We only prove the last point, the others following from classical arguments. We have that, since $|T| \leq \mathbf{N}_{\max}$, the matrix ${}^t\Psi_T\Psi_T$ is invertible and:

$$\|P_{V_T}x\|_2^2 = {}^t x P_{V_T} x = {}^t x \frac{1}{\sqrt{n}}\Psi_T\Big(\frac{1}{n}{}^t\Psi_T\Psi_T\Big)^{-1}\frac{1}{\sqrt{n}}{}^t\Psi_T x.$$

Since the eigenvalues of $\left(\frac{1}{n}{}^t\Psi_T\Psi_T\right)^{-1}$ are contained in $\left[\frac{1}{1+\nu}, \frac{1}{1-\nu}\right]$ we have proved that for all $x \in \mathbb{R}^p$:

$$\frac{1}{1+\nu}\frac{1}{n}\|{}^t\Psi_T x\|_2^2 \leq \|P_{V_T}x\|_2^2 \leq \frac{1}{1-\nu}\frac{1}{n}\|{}^t\Psi_T x\|_2^2.$$

7.2 Proof of lemma 2.2

Let I be a subset of $\{1, \ldots, p\}$, and define the following Gram matrix restricted to I:

$$G(I) := \frac{1}{n}{}^t\Psi_I\Psi_I.$$

Then, thanks to condition eq. (2.2), $G(I)$ has ones on its diagonal. Furthermore it is well known that all the eigenvalues of $G(I)$, which are real non negative, are included in disks centered at 1 and of radiuses:

$$\forall\, k \in I, \quad r_k = \sum_{j \in I\setminus\{k\}} \frac{1}{n}|<\psi_k, \psi_j>|.$$

And from the definition of the coherence those radiuses are uniformly bounded:

$$\forall\, k \in I, \quad |r_k| \leq (|I| - 1)\tau_n,$$

so that $|I| \leq \mathbf{N}_{\max}$ as soon as $(|I| - 1)\tau_n \leq \nu$, which shows that $\lfloor \nu/\tau_n \rfloor + 1 \leq \mathbf{N}_{\max}$.

7.3 Proof of proposition 2.6

Let S be a subset of indices such that $|S| \leq \mathbf{N}_{\max}$ and $I \subset S$. We start with the decomposition:

$$\|\alpha_S - \hat{\alpha}(S)\|_2^2 \leq 2(\underbrace{\|\alpha_S - \bar{\alpha}(S)\|_2^2}_{\text{Bias term}} + \underbrace{\|\bar{\alpha}(S) - \hat{\alpha}(S)\|_2^2}_{\text{Variance term}}).$$

The next two lemmas bound each of these two terms.

Lemma 2.12. *Let S be a subset of indices such that $|S| \leq \mathbf{N}_{\max}$. Then:*

$$\|\alpha_S - \bar{\alpha}(S)\|_2 \leq \frac{1}{1-\nu}\sqrt{|S|}\tau_n\|\alpha_{S^c}\|_1. \tag{2.23}$$

Proof. Indeed we know that for any subset of indices S, such that $|S| \leq \mathbf{N}_{\max}$, the matrix ${}^t\Psi_S\Psi_S$ is invertible, so that we can write:

$$\bar{\alpha}(S) = \alpha_S + ({}^t\Psi_S\Psi_S)^{-1}\ {}^t\Psi_S\Psi_{S^c}\alpha_{S^c},$$

from what we deduce that, using lemma 2.1:

$$\begin{aligned}\|\alpha_S - \bar{\alpha}(S)\|_2 &= \|({}^t\Psi_S\Psi_S)^{-1}\ {}^t\Psi_S\Psi_{S^c}\alpha_{S^c}\|_2,\\ &= \left\|(\frac{1}{n}{}^t\Psi_S\Psi_S)^{-1}(\frac{1}{n}{}^t\Psi_S\Psi_{S^c})\alpha_{S^c}\right\|_2,\\ &\leq \frac{1}{1-\nu}\left\|(\frac{1}{n}{}^t\Psi_S\Psi_{S^c})\alpha_{S^c}\right\|_2.\end{aligned}$$

If we denote by $\alpha_{S^c,i}$, $i \in S^c$, the components of α_{S^c} we have:

$$\left\|\frac{1}{n}{}^t\Psi_S\Psi_{S^c}\alpha_{S^c}\right\|_2 \leq \sum_{i\in S^c}|\alpha_{S^c,i}|\left\|\frac{1}{n}{}^t\Psi_S\psi_i\right\|_2,$$

and by definition of the coherence:

$$\forall i \in S^c, \quad \left\|\frac{1}{n}{}^t\Psi_S\psi_i\right\|_2 \leq \sqrt{|S|}\left\|\frac{1}{n}{}^t\Psi_S\psi_i\right\|_\infty \leq \sqrt{|S|}\tau_n.$$

□

Lemma 2.13. *Let S be a subset of indices such that $|S| \leq \mathbf{N}_{\max}$. Then:*

$$\|\hat{\alpha}(S) - \bar{\alpha}(S)\|_2^2 \leq \frac{1}{1-\nu}\frac{1}{n}\|P_{V_S}[\eta]\|_2^2. \tag{2.24}$$

Proof. Using lemma 2.1:

$$\begin{aligned}\|\bar{\alpha}(S) - \hat{\alpha}(S)\|_2^2 &\leq \frac{1}{1-\nu}\frac{1}{n}\|\Psi_S(\bar{\alpha}(S) - \hat{\alpha}(S))\|_2^2\\ &\leq \frac{1}{1-\nu}\frac{1}{n}\|P_{V_S}[\eta]\|_2^2.\end{aligned}$$

□

7.4 Proof of proposition 2.7

Let S be a subset of indices such that $|S| \leq \mathbf{N}_{\max}$ and $I \subset S$. We start with the decomposition:

$$\|\alpha_I - \hat{\alpha}(S)_I\|_2^2 \leq 3(\underbrace{\|\alpha_I - \bar{\alpha}(I)\|_2^2}_{t_1(I)} + \underbrace{\|\bar{\alpha}(I) - \hat{\alpha}(I)\|_2^2}_{t_2(I)} + \underbrace{\|\hat{\alpha}(I) - \hat{\alpha}(S)_I\|_2^2}_{t_3(I,S)}).$$

From lemma 2.12 we know that:

$$t_1(I) \leq \Big(\frac{1}{1-\nu}\Big)^2 |I| \tau_n^2 \|\alpha_{I^c}\|_1^2. \tag{2.25}$$

To bound t_2 we use lemma 2.13:

$$t_2(I) \leq \frac{1}{1-\nu}\frac{1}{n}\|P_{V_I}[\eta]\|_2^2.$$

Now to bound t_3 we first notice that, since $I \subset S$:

$$\begin{aligned}
\Psi_I\hat{\alpha}(I) - \Psi_I\hat{\alpha}(S)_I &= P_{V_I}(\Psi_I\hat{\alpha}(I) - \Psi_I\hat{\alpha}(S)_I), \\
&= P_{V_I}(\Psi_I\hat{\alpha}(I) - \Psi_S\hat{\alpha}(S) + \Psi_{S\setminus I}\hat{\alpha}(S)_{S\setminus I}), \\
&= P_{V_I}(P_{V_I}[Z] - P_{V_S}[Z] + \Psi_{S\setminus I}\hat{\alpha}(S)_{S\setminus I}), \\
&= P_{V_I}[\Psi_{S\setminus I}\hat{\alpha}(S)_{S\setminus I}].
\end{aligned}$$

So that using lemma 2.1 and the same reasoning as in the proof of lemma 2.12:

$$\begin{aligned}
t_3(I,S) &\leq \frac{1}{1-\nu}\frac{1}{n}\|\Psi_I(\hat{\alpha}(I) - \hat{\alpha}(S)_{|I})\|_2^2, \\
&= \frac{1}{1-\nu}\frac{1}{n}\|P_{V_I}[\Psi_{S\setminus I}\hat{\alpha}(S)_{|S\setminus I}]\|_2^2, \\
&\leq \frac{1}{(1-\nu)^2}\Big\|\frac{1}{n}{}^t\Psi_I\Psi_{S\setminus I}\hat{\alpha}(S)_{|S\setminus I}\Big\|_2^2, \\
&\leq \frac{1}{(1-\nu)^2}|I|\tau_n^2\|\hat{\alpha}(S)_{S\setminus I}\|_1^2.
\end{aligned}$$

Furthermore:

$$\begin{aligned}
\|\hat{\alpha}(S)_{S\setminus I}\|_1^2 &\leq 3(\|\bar{\alpha}(S)_{S\setminus I} - \alpha_{S\setminus I}\|_1^2 + \|\hat{\alpha}(S)_{S\setminus I} - \bar{\alpha}(S)_{S\setminus I}\|_1^2 + \|\alpha_{S\setminus I}\|_1^2), \\
&\leq 3(|S|t_1(S) + |S|t_2(S) + \|\alpha_{I^c}\|_1^2).
\end{aligned}$$

Now using that $I \subset S$:

$$|S|t_1(S) \leq \Big(\frac{1}{1-\nu}\Big)^2 |S|^2\tau_n^2\|\alpha_{S^c}\|_1^2 \leq \Big(\frac{1}{1-\nu}\Big)^2 (|S|\tau_n)^2\|\alpha_{I^c}\|_1^2,$$

and as we already showed:

$$|S|t_2(S) \leq \frac{|S|}{1-\nu}\frac{1}{n}\|P_{V_S}[\eta]\|_2^2.$$

Finally:

$$t_3(I,S) \leq \frac{3}{(1-\nu)^2}|I|\tau_n^2\Big\{\Big[\frac{1}{(1-\nu)^2}(|S|\tau_n)^2+1\Big]\|\alpha_{I^c}\|_1^2 + \frac{|S|}{1-\nu}\frac{1}{n}\|P_{V_S}[\eta]\|_2^2\Big\}.$$

And adding the three bounds we get:

$$\begin{aligned}\|\alpha_I - \hat{\alpha}(S)_I\|_2^2 \leq\ & |I|\tau_n^2\|\alpha_{I^c}\|_1^2\Big\{\frac{4}{(1-\nu)^2}+\frac{3}{(1-\nu)^4}(|S|\tau_n)^2\Big\}\\ &+\frac{1}{1-\nu}\frac{1}{n}\|P_{V_I}[\eta]\|_2^2 + \frac{3}{(1-\nu)^3}|I|\,|S|\tau_n^2\,\frac{1}{n}\|P_{V_S}[\eta]\|_2^2.\end{aligned}$$

7.5 Proof of proposition 2.8

We split the proof in small steps to make it easier to read.

— Let $I \subset \{1,\dots,p\}$ be a subset of indices such that $|I| \leq \mathbf{N}_{\max}$. Then the Gram matrix $\frac{1}{n}{}^t\Psi_I\Psi_I$ is invertible, and using lemma 2.1:

$$\|P_{V_I}[\eta]\|_2^2 \leq \frac{1}{1-\nu}\frac{1}{n}\|{}^t\Psi_I\eta\|_2^2. \tag{2.26}$$

— Since ${}^t\Psi_I\eta \sim \mathcal{N}(0, {}^t\Psi_I\Gamma\Psi_I)$, we immediately get that if I is a deterministic subset of indices, such that $|I| \leq \mathbf{N}_{\max}$, then:

$$\begin{aligned}\mathbb{E}\Big[\|P_{V_I}[\eta]\|_2^2\Big] &\leq \frac{1}{1-\nu}\frac{1}{n}\mathbb{E}\Big[\|{}^t\Psi_I\eta\|_2^2\Big],\\ &\leq \frac{1}{1-\nu}\frac{1}{n}\sum_{l\in I}\|\psi_l\|_\Gamma^2 := \frac{1}{1-\nu}\sigma_{\text{tot}}^2(I),\end{aligned}$$

where $\sigma_{\text{tot}}^2(I) = \frac{1}{n}\sum_{l\in I}\|\psi_l\|_\Gamma^2$. This proves the first inequality for deterministic subsets of indices.

— Let $c_\nu = \frac{1}{1-\nu}$, and for all indices $1 \leq l \leq p$, $u_l =< \psi_l, \eta > \sim \mathcal{N}(0, \|\psi_l\|_\Gamma^2)$. Then we can rewrite eq. (2.26) as:

$$\|P_{V_I}[\eta]\|_2^2 \leq c_\nu\frac{1}{n}\Big[\sum_{l\in I}u_l^2\Big].$$

Now define $S^{|I|-1} = \{x \in \mathbb{R}^{|I|}, \|x\|_2 = 1\}$ the unit sphere of $\mathbb{R}^{|I|}$ and

$$Z = \sqrt{\frac{1}{n}\sum_{l\in I}u_l^2} = \frac{1}{\sqrt{n}}\sup_{g\in S^{|I|-1}}\sum_{l\in I}g_l u_l = \frac{1}{\sqrt{n}}\sup_{g\in S^{|I|-1}}X_g,$$

where $X_g = \sum_{l\in I}g_l u_l$ is a centered Gaussian process indexed by the sphere $S^{|I|-1}$, and we can finally rewrite eq. (2.26) as $\|P_{V_I}[\eta]\|_2^2 \leq c_\nu Z^2$.

— We recall the following result from [69]: if X_t is a centered Gaussian process such that $\sigma^2 := \sup_t \mathbb{E}X_t^2$, then

$$\forall y > 0, \quad \mathbb{P}(\sup_t X_t - \mathbb{E}\sup_t X_t \geq y) \leq e^{-y^2/2\sigma^2}. \tag{2.27}$$

This is our main tool in handling concentration properties of suprema of centered Gaussian processes, and the last point proved that the random variable $\|P_{V_I}[\eta]\|_2^2$ is dominated by the square of such a supremum.

— First of all we bound the expectation of Z as:

$$\mathbb{E}\Big[\sup_{g\in S^{|I|-1}}\frac{1}{\sqrt{n}}X_g\Big]=\mathbb{E}[Z]\leq\sqrt{\mathbb{E}[Z^2]}=\sqrt{\frac{1}{n}\sum_{l\in I}\|\psi_l\|_\Gamma^2}:=\sigma_{\rm tot}(I). \tag{2.28}$$

Which implies that:

$$\forall\lambda\geq 0,\quad \mathbb{P}(Z^2\geq\lambda)\leq\mathbb{P}(Z-\mathbb{E}[Z]\geq\sqrt{\lambda}-\sigma_{\rm tot}(I)),$$

i.e.

$$\forall\lambda\geq 4\,\sigma_{\rm tot}^2(I),\quad \mathbb{P}(Z^2\geq\lambda)\leq\mathbb{P}(Z-\mathbb{E}[Z]\geq\sqrt{\lambda}/2). \tag{2.29}$$

— Then we seek a uniform bound on $\mathbb{E}[\frac{1}{n}X_g^2]$, for all $g\in S^{|I|-1}$. We have:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}X_g^2\Big]&=\frac{1}{n}\Big(\sum_{l\in I}g_l^2\|\psi_l\|_\Gamma^2+\sum_{l\neq l'}g_lg_{l'}{\rm Cov}(u_l,u_{l'})\Big),\\&=\sum_{l\in I}g_l^2\frac{\|\psi_l\|_\Gamma^2}{n}+\sum_{l\neq l'}g_lg_{l'}\frac{<\psi_l,\Gamma\psi_{l'}>}{n}.\end{aligned}$$

Define the Γ-coherence as:

$$\tau_n(\Gamma)=\max_{l\neq l'}\frac{|<\psi_l,\Gamma\psi_{l'}>|}{\|\psi_l\|_\Gamma\|\psi_{l'}\|_\Gamma}.$$

Then:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}X_g^2\Big]&\leq\max_{l\in I}\frac{\|\psi_l\|_\Gamma^2}{n}+\tau_n(\Gamma)\frac{1}{n}\sum_{l\neq l'}g_lg_{l'}\|\psi_l\|_\Gamma\|\psi_{l'}\|_\Gamma,\\&\leq\max_{l\in I}\frac{\|\psi_l\|_\Gamma^2}{n}+\tau_n(\Gamma)\frac{1}{n}(\sum_{l\in I}g_l\|\psi_l\|_\Gamma)^2,\\&\leq\max_{l\in I}\frac{\|\psi_l\|_\Gamma^2}{n}+\tau_n(\Gamma)\frac{1}{n}\sum_{l\in I}\|\psi_l\|_\Gamma^2.\end{aligned}$$

We define for any subset of indices I:

$$\sigma_*^2(I):=\max_{l\in I}\frac{\|\psi_l\|_\Gamma^2}{n}+\tau_n(\Gamma)\sigma_{\rm tot}^2(I). \tag{2.30}$$

— Then from eq. (2.27) we get:

$$\forall y>0,\quad \mathbb{P}(Z-\mathbb{E}Z\geq y)\leq e^{-y^2/2\sigma_*^2(I)}.$$

So that, using eq. (2.29), for all $\lambda\geq 4\sigma_{\rm tot}^2(I)$ we have:

$$\mathbb{P}(Z^2\geq\lambda)\leq\mathbb{P}(Z-\mathbb{E}[Z]\geq\sqrt{\lambda}/2)\leq e^{-\lambda/8\sigma_*^2(I)}.$$

Finally, since $\|P_{V_I}[\eta]\|_2^2\leq c_\nu Z^2$, we have as soon as $\lambda\geq 4c_\nu\sigma_{\rm tot}^2(I)$ and $1\leq|I|\leq\mathbf{N}_{\max}$:

$$\mathbb{P}(\|P_{V_I}[\eta]\|_2^2\geq\lambda)\leq e^{-\lambda/8c_\nu\sigma_*^2(I)}. \tag{2.31}$$

— From eq. (2.31) we get for any deterministic subsets of indices I such that $|I| \leq \mathbf{N}_{\max}$:

$$\begin{aligned}
\mathbb{E}[\|P_{V_I}[\eta]\|_2^4] &= \mathbb{E}[\|P_{V_I}[\eta]\|_2^4 \mathbb{1}\{\|P_{V_I}[\eta]\|_2^4 \leq 16c_\nu^2\sigma_{\text{tot}}^4(I)\} \\
&\quad + \|P_{V_I}[\eta]\|_2^4 \mathbb{1}\{\|P_{V_I}[\eta]\|_2^4 \geq 16c_\nu^2\sigma_{\text{tot}}^4(I)\}], \\
&\leq 16c_\nu^2\sigma_{\text{tot}}^4(I) + \int_{16c_\nu^2\sigma_{\text{tot}}^4(I)}^{+\infty} \mathbb{P}(\|P_{V_I}[\eta]\|_2^4 \geq \lambda) d\lambda, \\
&\leq 16c_\nu^2\sigma_{\text{tot}}^4(I) + \int_{16c_\nu^2\sigma_{\text{tot}}^4(I)}^{+\infty} e^{-\sqrt{\lambda}/8c_\nu\sigma_*^2(I)} d\lambda, \\
&\leq 128c_\nu^2(\sigma_{\text{tot}}^2(I) + \sigma_*^2(I))^2.
\end{aligned}$$

Notice that $\sigma_*^2(I) \leq (1+\tau_n(\Gamma))\sigma_{\text{tot}}^2(I) \leq 2\sigma_{\text{tot}}^2(I)$, so that:

$$\mathbb{E}[\|P_{V_I}[\eta]\|_2^4] \leq 1152\, c_\nu^2\sigma_{\text{tot}}^4(I), \tag{2.32}$$

which proves the second inequality for deterministic subsets of indices.

— It is convenient from now on to introduce the quantity $\sigma^2(I)$, defined on any subset of indices I as:

$$\sigma^2(I) = \sum_{l\in I} \frac{\|\psi_l\|_\Gamma^2}{n} \vee 1 := \sum_{l\in I} \sigma_l^2,$$

so that for any subset I, we have at the same time $\sigma_{\text{tot}}^2(I) \leq \sigma^2(I)$ and $|I| \leq \sigma^2(I)$. Define $\mathbb{T}_{\Sigma_*}^N = \{I; \sigma^2(I) \leq \Sigma_*, |I| \leq N\}$. Let now $\tilde{I}$ be a random subset of $\mathbb{T}_{\Sigma_*}^N$, then by union bound:

$$\begin{aligned}
\forall \lambda \geq 4c_\nu\Sigma_*, \quad \mathbb{P}(\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq \lambda) &\leq \sum_{I\in\mathbb{T}_{\Sigma_*}^N} \mathbb{P}(\|P_{V_I}[\eta]\|_2^2 \geq \lambda), \\
&\leq p^N e^{-\lambda/16c_\nu\Sigma_*}, \\
&\leq \exp\Big(N\log p - \frac{\lambda}{16c_\nu\Sigma_*}\Big).
\end{aligned}$$

Then as soon as $p \geq 2$:

$$\forall \lambda \geq 32c_\nu\Sigma_* N\log p, \quad \mathbb{P}(\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq \lambda) \leq e^{-\lambda/32c_\nu\Sigma_*}. \tag{2.33}$$

— Now to bound $\mathbb{E}[\|P_{V_{\tilde{I}}}[\eta]\|_2^2]$, where $\tilde{I}$ is a random subset of $\mathbb{T}_{\Sigma_*}^N$, we observe that, as soon as $p \geq 2$:

$$\begin{aligned}
\mathbb{E}[\|P_{V_{\tilde{I}}}[\eta]\|_2^2] &= \mathbb{E}[\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \mathbb{1}\{\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \leq 32c_\nu\Sigma_* N\log p\} \\
&\quad + \|P_{V_{\tilde{I}}}[\eta]\|_2^2 \mathbb{1}\{\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq 32c_\nu\Sigma_* N\log p\}], \\
&\leq 32c_\nu\Sigma_* N\log p + \int_{32c_\nu\Sigma_* N\log p}^{+\infty} \mathbb{P}(\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq \lambda) d\lambda, \\
&\leq 32c_\nu\Sigma_* N\log p + \int_{32c_\nu\Sigma_* N\log p}^{+\infty} e^{-\lambda/32c_\nu\Sigma_*} d\lambda, \\
&\leq 64c_\nu\Sigma_* N\log p.
\end{aligned}$$

And in the same way:

$$\mathbb{E}[\|P_{V_{\tilde{I}}}[\eta]\|_2^4] \leq 5120 c_\nu^2 \Sigma_*^2 N^2 (\log p)^2. \tag{2.34}$$

— Let $\sigma^2_{\max}$ and Σ_* be deterministic quantities, and define:

$$\mathbb{M}_{\sigma^2_{\max},\Sigma_*,N} = \Big\{ I;\, \max_{l\in I} \frac{\|\psi_l\|_\Gamma^2}{n} \leq \sigma^2_{\max}, \sigma^2(I) \leq \Sigma_*, |I| \leq N \Big\}.$$

Suppose that $\tilde{I}$ is a random subset of $\mathbb{M}_{\sigma^2_{\max},\Sigma_*,N}$. Then by union bound:

$$\begin{aligned}
\forall \lambda \geq 4c_\nu \Sigma_*, \quad \mathbb{P}(\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq \lambda) &\leq \sum_{I \in \mathbb{M}_{\sigma^2_{\max},\Sigma_*,N}} \mathbb{P}(\|P_{V_I}[\eta]\|_2^2 \geq \lambda), \\
&\leq p^N e^{-\lambda/8c_\nu(\sigma^2_{\max} + \tau_n(\Gamma)\Sigma_*)}, \\
&\leq \exp\Big(N \log p - \frac{\lambda}{8c_\nu(\sigma^2_{\max} + \tau_n(\Gamma)\Sigma_*)} \Big).
\end{aligned}$$

Then:

$$\forall \lambda \geq 16 c_\nu (\sigma^2_{\max} + \tau_n(\Gamma)\Sigma_*) N \log p, \quad \mathbb{P}(\|P_{V_{\tilde{I}}}[\eta]\|_2^2 \geq \lambda) \leq e^{-\lambda/16c_\nu(\sigma^2_{\max} + \tau_n(\Gamma)\Sigma_*)}.$$

This implies by the same reasoning as before that:

$$\mathbb{E}[\|P_{V_{\tilde{I}}}[\eta]\|_2^2] \leq 32 c_\nu (\sigma^2_{\max} + \tau_n(\Gamma)\Sigma_*) N \log p.$$

— Define, for any subset of indices I, the quantity $\sigma^2_{\max}(I) = \max_{l\in I} \sigma_l^2$. Let L be a random subset of indices belonging to $L_{\lambda,\Sigma_*,N}$. Finally let $S = S_{\frac{\lambda}{2}\sigma}(\alpha)$ for some $\alpha \in \mathbb{R}^p$.
Then:

$$\begin{aligned}
\mathbb{E}[\|P_{V_L}[\eta]\|_2^2] = \; & \underbrace{\mathbb{E}\Big[\|P_{V_L}[\eta]\|_2^2 \mathbb{1}\Big\{ \sigma^2_{\max}(L) \leq \sigma^2_{\max}(S) \Big\}\Big]}_{A} \\
& + \underbrace{\mathbb{E}\Big[\|P_{V_L}[\eta]\|_2^2 \mathbb{1}\Big\{ \sigma^2_{\max}(L) > \sigma^2_{\max}(S) \Big\}\Big]}_{B}.
\end{aligned}$$

On the event $\Big\{ \sigma^2_{\max}(L) \leq \sigma^2_{\max}(S) \Big\}$, L is a random subset of $\mathbb{M}_{\sigma^2_{\max}(S),\Sigma_*}$, which implies, by the reasoning of the preceding point, that:

$$A \leq 32 c_\nu (\sigma^2_{\max}(S) + \tau_n(\Gamma)\Sigma_*) N \log p.$$

It remains to handle B. To do so we will bound:

$$\mathbb{P}(\sigma^2_{\max}(L) > \sigma^2_{\max}(S)).$$

This probability is inferior to the probability of having any index external to S among L. And the probability that a given index l belongs to L and not to S can

be bounded in the following manner:

$$\begin{aligned}\mathbb{P}(l \in L, l \notin S) &\leq \mathbb{P}(|\check{\alpha}_l/\sigma_l| \geq \lambda, |\alpha_l/\sigma_l| < \lambda/2) \\ &\leq \mathbb{P}(|\check{\alpha}_l - \alpha_l| \geq \sigma_l\lambda/2), \\ &\leq 2e^{-n\lambda^2/16},\end{aligned}$$

for every $\lambda \geq 8\tau_n\|\alpha\|_1/\sigma_l$. Then by union bound:

$$\mathbb{P}(\sigma^2_{\max}(L) > \sigma^2_{\max}(S)) \leq 2pe^{-n\lambda^2/16},$$

as soon as $\lambda \geq 8\tau_n\|\alpha\|_1$ (since for all l, $\sigma_l \geq 1$). Now using the bound eq. (2.34) and Cauchy-Schwarz inequality:

$$\begin{aligned}B &\leq \sqrt{\mathbb{E}[\|P_{V_L}[\eta]\|_2^4]}\sqrt{\mathbb{P}(\sigma^2_{\max}(L) > \sigma^2_{\max}(S))} \\ &\leq 102c_\nu\Sigma_* N\log(p)\sqrt{p}e^{-n\lambda^2/32},\end{aligned}$$

for all $\lambda \geq 8\tau_n\|\alpha\|_1$.

Then we have $B \leq A$ as soon as:

$$\lambda^2 \geq (8\tau_n\|\alpha\|_1)^2 \vee \Big(32\frac{1}{n}\log\Big(\frac{102}{32}\frac{\Sigma_*\sqrt{p}}{\sigma^2_{\max}(S) + \tau_n(\Gamma)\Sigma_*}\Big)\Big). \tag{2.35}$$

Finally using that $\sigma^2_{\max}(S_{\lambda/2,\sigma}) \geq 1$ and since $\Sigma_* \leq p^\theta$, for some $\theta > 0$, we get that, for $p \geq 2$,

$$\mathbb{E}[\|P_{V_L}[\eta]\|_2^2] \leq 64c_\nu(\sigma^2_{\max}(S) + \tau_n(\Gamma)\Sigma_*)N\log p, \tag{2.36}$$

as soon as $\lambda^2 \geq 144\,\theta\Big[(\tau_n\|\alpha\|_1)^2 \vee \frac{\log(p)}{n}\Big]$.

Chapter 3

Orthogonal matching pursuit with pivoting: accelerating greedy pursuit algorithms

This chapter is the replica of an article submitted to a scientific review. It can be read independently from the rest of the manuscript.

Abstract

Model selection is the focus of a lot of attention in contemporary statistics, especially for high dimensional models where the number of covariates, p, is greater than the number of observations, n. Among the most common approach to linear regression in a high dimensional setting, greedy methods are known to perform well. But on the other hand, while fast they still require frequent calls to a scoring procedures, in fact each time they want to incorporate a new covariate. To avoid this computational burden super greedy methods, where a whole bunch of covariates are incorporated at each step, have been developed. This paper introduces a super greedy modification of orthogonal matching pursuit (OMP) with a "pivoting" rule which tries to get the best of both world: OMP-like efficiency while still making as few call to the scoring process as possible. The result is procedure much faster than OMP with comparable predictive efficiency. Furthermore we demonstrate on real data that the ability of OMP with pivoting to select correlated covariates may allow to reach an even better prediction accuracy than orthogonal matching pursuit.

Contents

1 Introduction

The question of model selection is central to modern statistics. Consider linear models, where the go-to tool is the least squares estimation. Many situations occur where the statistician should not use all the covariates at his disposal to compute the least squares estimator. Indeed, it is common nowadays to face data sets where the number of available covariates, p, may exceed the number of available observations, n. In such a situation, the so-called high dimensional case, least squares estimation may behave very poorly. Even for "classical" situations, where $n > p$, it is well known [71] that every covariate that we decide to use increases the variance of the least squares estimator, while not necessarily diminishing its bias in a comparable proportion. These considerations lead us to the problem of sparse linear regression: can we provide an estimator with good approximation properties, while still having many zero coefficients ?

More precisely we are concerned by linear models of the form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{3.1}$$

where $\mathbf{y}_{n\times 1}$ is the response vector, $\mathbf{X}_{n\times p}$ is the design matrix. The parameter of the model is $\beta_{p\times 1}$ and $\varepsilon_{n\times 1}$ is an additive error. We consider both classical models, where $n \geq p$, and high dimensional ones, where $n < p$. A formulation of the sparse regression problem looks for the best possible estimator, $\hat{\beta}$, with a maximum sparsity of N:

$$\begin{cases} \min_{\hat{\mathbf{y}}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2 = \min_{\hat{\beta}} \|\mathbf{y} - X\hat{\beta}\|_2, \\ \text{s.t. } \|\hat{\beta}\|_0 \leq N. \end{cases} \tag{3.2}$$

Exhaustive search among all possible subsets of N elements, usually referred to as best subset selection, is already intractable when p is moderately large. To provide a computationally efficient solution, two main approaches have been explored. The first one is to relax the ℓ_0 penalty to a convex one, where the choice of the penalty ℓ_1 is usually referred to as the Lasso [89], the choice ℓ_2 as the ridge regression, and the in between as the elastic net regression [105]. Another approach to the computation of a solution to eq. (3.2) is the greedy approximation heuristic [88], where covariates are incorporated one-by-one, each step resulting from a locally optimal choice.

This paper focuses on the orthogonal matching pursuit algorithm introduced in [79], [30] and its super greedy acceleration.

1.1 Orthogonal Matching Pursuit

One of the most straightforward greedy approaches to the sparse regression problem is the so-called Orthogonal Matching Pursuit(OMP) procedure. It incorporates covariates one-by-one, each time selecting the covariate which is the most correlated to the residual vector left from the preceding step. Once the covariate is selected, the observed response $\mathbf{y}$

is orthogonally projected onto the span of the selected covariates, the residual is recomputed, and the process repeats. We will call pivoting the fact of recomputing the residuals and performing a new sensing step. More precisely:

1. Start from an initial residual vector $r^0 = \mathbf{y}$, an initial parameter approximation $\hat{\beta}^0 = 0$, and an initial set of selected variables $S^0 = \emptyset$.
2. Suppose that you have constructed r^{n-1}, β^{n-1} and S^{n-1} without having met some stopping criterion. Then:
   — compute the inner products $< x_i, r^{n-1} >$ for all $i \notin S^{n-1}$, where the x_i's are the column vectors of the design $\mathbf{X}$. This is the **scoring** step.
   — let $i_{\max} = \arg\max_{i \notin S^{n-1}} | < x_i, r^{n-1} > |$.
   — update the selected indices $S^n = S^{n-1} \cup \{i_{\max}\}$.
   — compute the least squares solution

$$\hat{\beta}^n = \arg\min_{\beta} \|Y - X\beta\|_2.$$

   — update the residuals, $r^n = \mathbf{y} - \mathbf{X}\hat{\beta}^n$.

This procedure enjoys strong theoretical guarantees, when the design is well behaved. Indeed it is known that N steps of OMP provide us with an estimator not too far from the optimal N sparse estimator, as long as the design is "almost" orthogonal. This almost orthogonality notion is usually quantified in one of two ways: we can require the coherence defined as

$$\mu = \max_{i \neq j} | < x_i, x_j > |. \tag{3.3}$$

(or a refinement of it, like the cumulative coherence) to be small [52], [90] relatively to the sparsity of the signal (i.e. the less sparse the signal, the lower the coherence), or we can work under the Restricted Isometry Property (RIP) [20] which requires all sub-matrices of the design extracted from picking k columns to the original design to behave almost as isometries. and still obtain good results [102]. Those hypotheses are not too restrictive when we are allowed to choose the design matrix (like in compress sensing) but can be very restrictive from a statistician perspective who has usually no control over the design.

The stopping criterium comes essentially in two flavors: you can choose the number of steps a priori, i.e. fix a target sparsity, or stop when the residuals norm, $\|r^n\|_2$, gets smaller than some target error $\epsilon > 0$. It is important to notice that the stopping criterium plays the role of a regularization parameter and some theoretical results exist on its choice. For example in the case of Gaussian noise with standard deviation σ, the stopping condition $\|{}^t\mathbf{X}r^n\|_\infty \leq C\sigma\sqrt{\log(p)}$, for some constant C, is known to perform well when the non zero components of the coefficient vector are large enough [16].

Furthermore OMP is very efficient from a computational point of view, if implemented using a QR or Cholesky process. In algorithm 1 we describe the QR process implementation.

Such a progressive Gram-Schmidt procedure allows the computation of $\hat{\beta}^n$ to rely on the computations of the preceding step, instead of starting from scratch at each step.

Algorithm 1 OMP with QR process

Input: observed response $\mathbf{y}$, design $\mathbf{X}$, target sparsity K or target error ε

Output: estimated parameter $\hat{\beta}$

Init: Set $S^0 = \emptyset$, $r^0 = \mathbf{y}$, $\hat{\beta}^0 = 0$, $Q_0 = 0$, $R_0 = 0$

while Stopping criterion not met **do**

$\hat{i} = \arg\max_i | < x_i, r^{k-1} > |$

$v \leftarrow x_{\hat{i}}$

$S^k \leftarrow S^{k-1} \cup \{\hat{i}\}$

$R_k \leftarrow \begin{pmatrix} R_{k-1} & R_{\bullet k} \\ 0 & \end{pmatrix}$ ▷ $R_{\bullet k}$ is the new k-th column of R_k

$Q_k \leftarrow \begin{pmatrix} Q_{k-1} & q_k \end{pmatrix}$

for i = 1:(k - 1) **do**

$R_{ik} \leftarrow < q_i, v >$

$v \leftarrow v - R_{ik} q_i$

end for

$R_{kk} \leftarrow \|v\|_2$

$q_k \leftarrow v / R_{kk}$

$k \leftarrow k + 1$

end while

▷ We denote by Q and R the resulting matrices of the while loop

$s \leftarrow {}^tQ\mathbf{y}$

Solve $s = R\hat{\beta}$ ▷ Back Substitution

Suppose that we perform N iterations (and output a N-sparse estimator) of OMP. At each step it is necessary to:

— compute the inner products $< x_i, r^{k-1} >$, each inner product is at most a linear computation in n, for a total of $O(Nnp)$ computations (finding the maximum is linear in p)

— at step k, filling the vector $R_{\bullet k}$ is essentially a $O(kn)$ computation, and is the most demanding operation of the QR process. Summing over k, N steps of the QR process require $O(N^2 n)$ computations

— the final back substitution requires $O(N^2)$ computations.

In the end, N steps of OMP with a QR process will require $O(Nnp + N^2 n + N^2)$ computations. It is important to notice that for high-dimensional models it is the scoring process, which scales linearly with the number of covariates, which is going to dominate the total cost of the method.

It can be noticed that with little additional cost, the QR process we described allows us to compute the whole path of OMP solutions for sparsity varying between 1 and N. Indeed, for any $1 \leq k \leq N$, the k first columns of Q and the $k \times k$ upper-left sub-matrix of R are the result of k steps OMP, so that only the last back-substitution has to be iterated. Then the total cost is $O(Nnp + N^2n + N^3)$ for the computation of the whole path of OMP estimators, for sparsities varying between 1 and N. Since necessarily $N \leq n$, the pathwise modification of OMP which computes all the solutions with at most N steps requires a total of $O(Nnp + N^2n)$ computations. This makes the efficient selection of the stopping step by cross-validation possible.

1.2 Super Greedy modification of OMP

A second look at OMP reveals an important computational bottleneck. Indeed at each iteration we select a covariate. This selection process can be split into two main computations:

1. each covariates is given a score. At iteration n covariates x_i gets the score

$$s_i = | < r^n, x_i > |,$$

2. sorting those scores induces an order on the design columns' indices, which can be encoded in a permutation π_s. Then $x_{\pi_s(1)}$ is the column of the design the most correlated (in absolute value) to the residuals, and $x_{\pi_s(p)}$ is the least correlated column.

It is interesting at this point to consider OMP behavior when the design $\mathbf{X}$ is orthogonal. The initial correlation vector, $c^0 = {}^t\mathbf{X}\mathbf{y}$, specifies which covariate, x_{π_1}, enters the set of selected covariates first. The new residual vector is then $r^1 = \mathbf{y} - < \mathbf{y}, q_1 > q_1$. Then if we update the correlation vector to $c^1 = {}^t\mathbf{X}r_1$, we find the same scores for all the covariates but x_{π_1}. It is then unnecessary to update the correlation vector, and so to recompute the correlations !

We may hope for this idea to still be usable in the case where the design is not too correlated: do not update the scores at each iteration, rather let not only the most correlated covariate enter the set of selected ones at each iteration but say the $k > 1$ most correlated before updating the scores. Doing so, for a target sparsity of N we only perform the scoring step N/k times. This idea, termed super greedy in [67], [66], aims at achieving the same efficiency as OMP but with much less scoring steps, and so with much less computations. The same idea is already studied in [43] where each step of the greedy procedure incorporates all the covariates with a score higher than some threshold. The thresholding methodology relies on the notion of False Discovery Rate, [1]. Finally an extreme point of view is developed in [63], [73] by analogy with the orthogonal case (where OMP is particularly inefficient), since only one scoring operation is performed and all covariates with a score above some threshold are included at step one. We will refer

to this extreme procedure as one-step OMP. All those strategies are backed by strong theoretical results under the same kind of hypothesis than OMP.

The question is then to find the right balance between the number of iterations of the procedure and the size of the steps. This is illustrated in fig. 3.1, where we can see that if the steps are small (2 or 4 atoms by step here) the super greedy variation mimics OMP behavior very well but the computational cost is still high. On the other hand for very large steps (here 40 or even 80 atoms by step) the computational cost is drastically reduced by the procedure performs badly. Indeed its error tends to plateau, since irrelevant covariates are included, which could be avoided by updating the their scores.

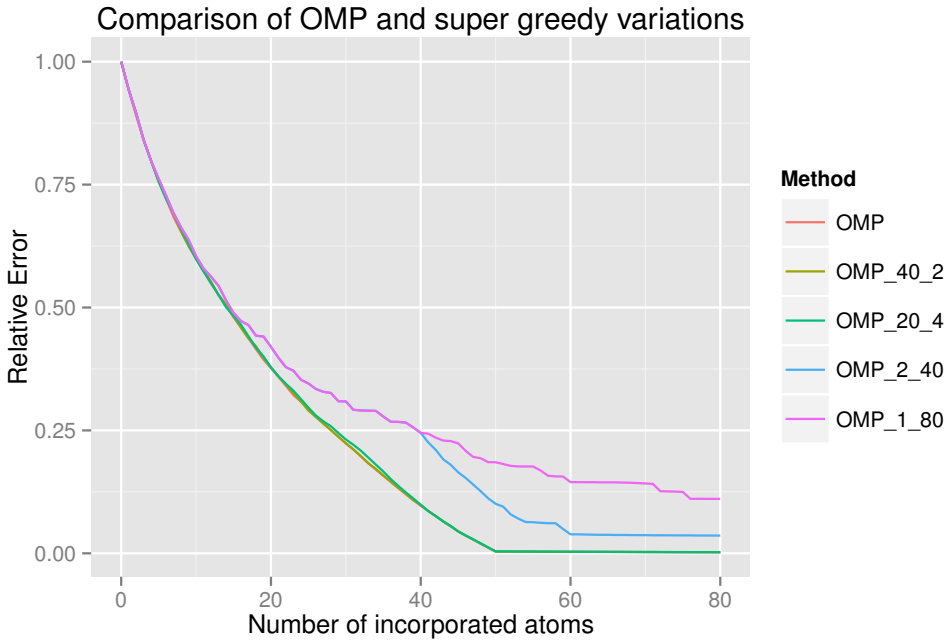

FIGURE 3.1. Comparison of OMP with different super greedy variations. The design is a random Gaussian design with $n = 500$ and $p = 1500$. The parameter β is sparse, with sparsity $S = 50$. We use the following notation: OMP_N_k means N iterations of OMP with steps of size k.

This paper aims at designing a method with an efficiency comparable to that of OMP, while still making as few calls as possible to the scoring procedure, by introducing a pivoting rule which adaptively decides when to update the scores.

2 Super Greedy OMP with pivoting rule

We want to achieve comparable prediction performance to OMP while still making as few calls to the scoring procedure as possible. To do so we consider a "pivoting" rule P. Starting from an initial score vector, s, we incorporate covariates one-by-one in the order induced by s as long as P is verified. Only when a new covariate does not pass P, do we update the scores and proceed to the next greedy step. Of course a pivoting rule P can only have access to the information produced by previous steps of the procedure. Since

the drawback of making very long steps is the tendency of the error to plateau, we will use the norm of the current residuals, compared to the norm of the residuals at the step before, as our decision criterium.

The super greedy OMP procedure with pivoting is described as follows, given an extra parameter $\lambda \in (0,1)$:

1. Start from an initial residual vector $r^0 = \mathbf{y}$, an initial parameter approximation $\hat{\beta}^0 = 0$, an initial set of selected variables $S^0 = \emptyset$ and an initial score vector $s = [\,| < x_1, \mathbf{y} > |, \ldots, | < x_p, \mathbf{y} > |\,]$.
2. incorporate covariates one-by-one in the order of their scores. Each time a new covariate is incorporated update the QR decomposition as in algorithm 1, to form Q_k and R_k (at step k).
3. after each update of the QR decomposition compute the new residuals $r^k = \mathbf{y} - {}^tQ_k\mathbf{y}$.
4. compute the ratio $R = \|r^k\|_2 / \|r^{k-1}\|_2$.
5. Pivoting rule: if $R < \lambda$ proceed to the next covariate. If not restart the procedure from the $(k-1)$-th step, and update the scores before selecting the k-th covariate.

Following this super greedy procedure, we decide to incorporate covariates without updating the scores, as long as:

$$\frac{\|r^k\|_2}{\|r^{k-1}\|_2} < \lambda.$$

If we denote by q_i the columns of the QR process associated to OMP, for $1 \leq i \leq k$, we can rewrite our condition as:

$$< y, q_k >^2 > (1-\lambda^2)\Big\{ \|y\|_2^2 - \sum_{i=1}^{k-1} < y, q_i >^2 \Big\}. \tag{3.4}$$

This can be interpreted as asking that the energy added by the introduction of a new direction q_k, $< y, q_k >^2$, is higher than a constant, $1-\lambda^2$, times the missing energy from the preceding step, $\|y\|_2^2 - \sum_{i=1}^{k-1} < y, q_i >^2$. In practice we choose λ close to 1, if not the process tends to stop very soon. Of course the number of steps that we perform can be chosen by cross validation easily by the simple pathwise modification already described for OMP. In the same way we can optimize too on the extra λ parameter by cross validation.

3 Numerical Studies

In this section we will demonstrate the efficiency of our method by numerical experiments. We focus on two aspects: reaching comparable performance to that of OMP while having a much lower computational cost.

3.1 Simulation data

The design matrices $\mathbf{X}$ considered in this study are of random type and built on $n \times p$ independent and identically distributed normal standard random variables, a favorable

setting for OMP since they have an almost orthogonal behavior with high probability. Given $\mathbf{X}$, the target observations are $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. The vector of parameters β is simulated as follows: all coordinates are zero except S picked uniformly at random among the p possible choices. If the l-coordinate has been picked, we take $\beta_l = r_l|z_l|$, where r_l is a Rademacher random variable, i.e. takes the value $+1$ or -1 with equal probability, and $z_l \sim \mathcal{N}(5, 1)$. The vector β is then renormalized to fix its norm, so that the signal-to-noise ratio (SNR) is around 10. In each case we split the design and the observed vector between a train set with 75 % of the observations, and the 25 % remaining in a test set. We estimate the coefficient vector on the train set and evaluate its accuracy on the independent test set. We consider the normalized prediction error:

$$\frac{\|\mathbf{X}_{\text{test}}\beta - \mathbf{X}_{\text{test}}\hat{\beta}\|_2^2}{\|\mathbf{X}_{\text{test}}\beta\|_2^2}.$$

On fig. 3.2 we compare the two extremes, OMP and one-step OMP, with our method with an additional pivoting. We can see that at first our methodology is the same as one step OMP, but as soon as its training error begins to plateau, the scores are updated which allows the method to break from one-step OMP and get closer to OMP.

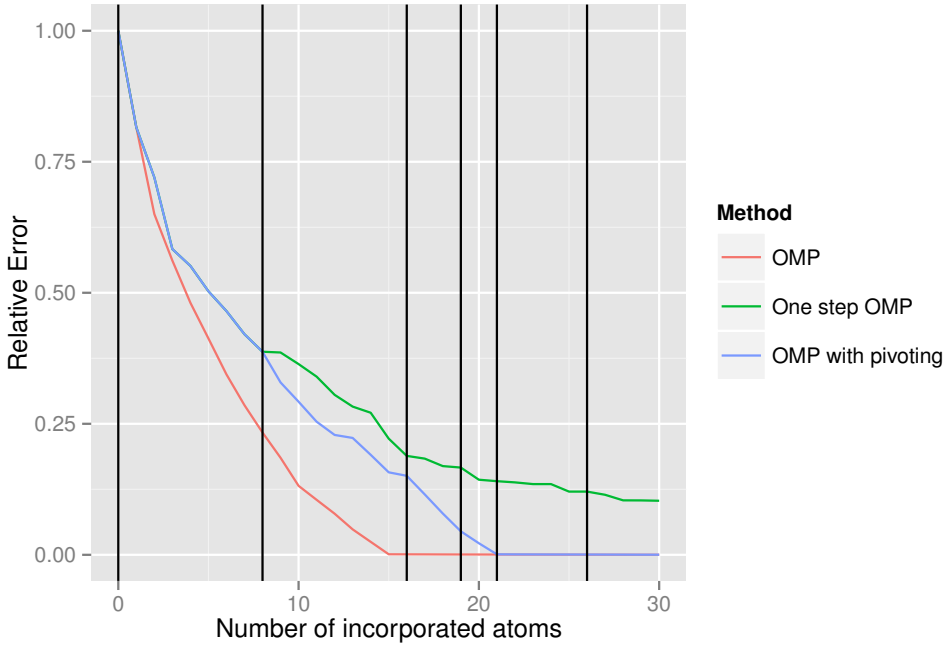

FIGURE 3.2. Comparison of OMP, one-step OMP and our methodology. The design is a random Gaussian design with $n = 75$ and $p = 300$. The parameter β is sparse, with sparsity $S = 15$. The vertical lines show when the pivoting updates the scores of the covariates.

We can compare the typical relative prediction error of the two methods on these random designs by repeating the experiment many times, and selecting for each method

an optimal number of steps by cross validation. The results are reported on table 3.1.

OMP	OMP with pivoting
Min. :0.003784	Min. :0.003784
1st Qu.:0.004677	1st Qu.:0.004757
Median :0.005245	Median :0.005490
Mean :0.005256	Mean :0.006436
3rd Qu.:0.005685	3rd Qu.:0.006210
Max. :0.007899	Max. :0.022864

TABLE 3.1. Summary of the relative prediction error for OMP and OMP with pivoting. The design is a random Gaussian design with $n = 750$ and $p = 1500$. The parameter β is sparse, with sparsity $S = 50$, and we make 100 repetitions.

A graphic summary of those results using box plots is reported on fig. 3.3.

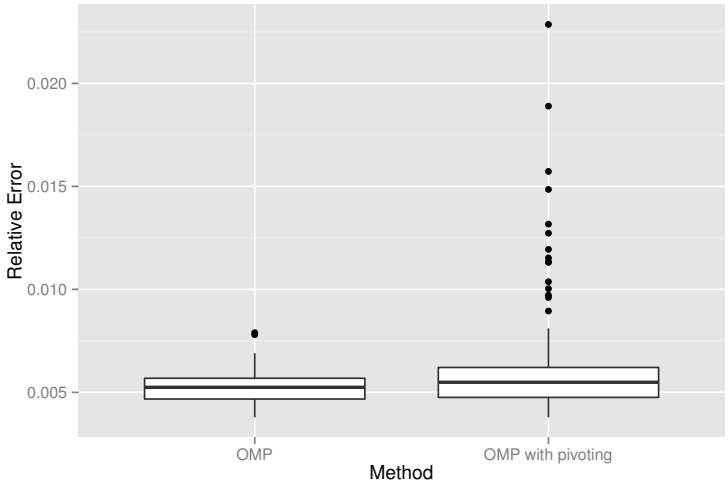

FIGURE 3.3. Comparison of OMP and our methodology. The design is a random Gaussian design with $n = 750$ and $p = 1500$. The parameter β is sparse, with sparsity $S = 50$, and we make 100 repetitions.

All these experiments show that OMP with pivoting has a comparable prediction efficiency to OMP. The great benefit of the pivoting rule comes from its computation time. Indeed OMP with pivoting is in general much faster than OMP (the number of pivots is not a deterministic quantity so that we can only assess the pivoting strategy superiority empirically). To do so we compare the execution time of the two procedures for designs with fixed n but with a growing number of covariates p. To make the comparison fair, both procedures are required to perform the same number of iterations, i.e. to produce estimators with the same sparsity (and as shown before with comparable prediction power). Results are reported on fig. 3.4. We can see that even if both methods seem to scale linearly as a function of the number of covariates, OMP with pivoting is a clear winner with much weaker dependency on the dimension p.

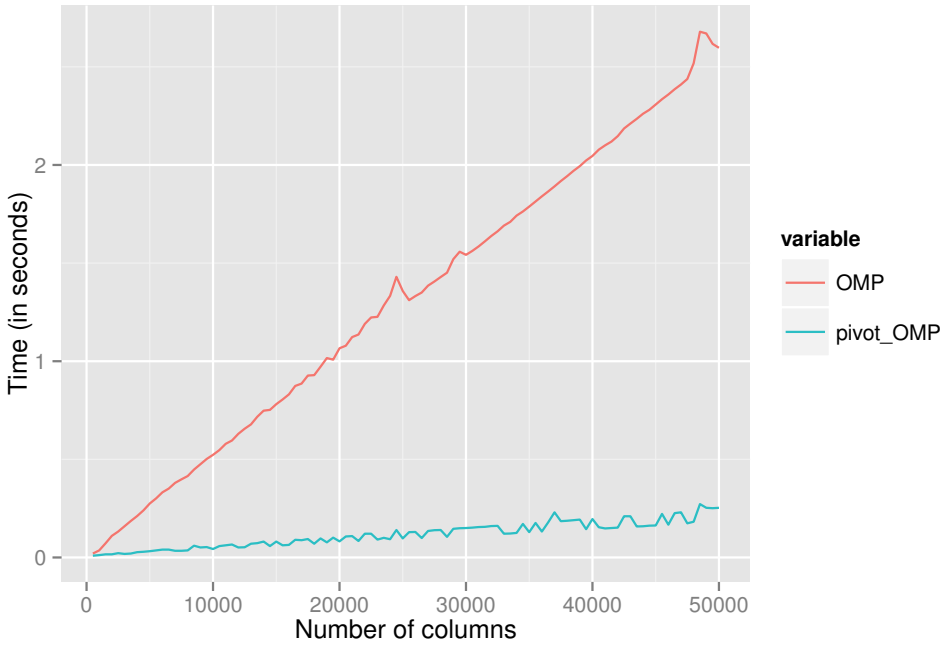

FIGURE 3.4. Comparison of the execution time of OMP and our methodology. The design is a random Gaussian design with fixed $n = 750$. The parameter β is sparse, with sparsity $S = 50$, and the two methods are iterated for 150 steps.

3.2 Real-world texts data sets

We consider a corpus of documents of interest $\{D_1, \ldots, D_N\}$. A classical model in natural language processing is the bag-of-words for which the texts are represented as a bag (multiset) of their words. To such a model it is natural to associate a matrix representation, the Document-Term Matrix representation, which is a matrix where the lines represent a document in the corpus, and the columns a word present among the documents. In this section we will consider the simplest encoding, the binary weighting. In this case the coefficient (i, j) of the Document-Term matrix M is equal to 1 if the word j is present in the document i. In this case each line of M is the realization of a multidimensional Bernoulli random variable. The resulting matrix is usually very high dimensional but at the same time very sparse, so that it can still be stored using specific sparse matrices format.

We will analyze here a dataset containing many jobs advertisements, with the obtained salary, which can be obtained from the Kaggle competition `https://www.kaggle.com/c/job-salary-prediction`. We try to predict the salary given only the text of the ad. We compare the selection operated by OMP on one side with the one of OMP with pivoting. We use 5000 ads which leads to a 5000×15955 matrix. We compare OMP and OMP with pivoting on this data set, plotting the relative prediction error of the two methods as a function of the number of iterations. The results are reported on fig. 3.5. In is interesting to observe that the pivoting methodology here clearly outperforms classical orthogonal matching pursuit.

FIGURE 3.5. Comparison of the relative prediction error of OMP and and OMP with pivoting on text data with binary encoding.

It is interesting to compare the first 10 terms selected by each method in table 3.2. Both methods select "and" at first (since it is a very frequent word, this artifact could be removed by first removing its mean from the salary vector, but we wanted the minimal amount of preprocessing, to not rely on natural processing techniques). OMP with pivoting updates the scores twice during the first ten steps which allows it to select correlated variables (like "projects" and "projects") and reach a better accuracy. On the other hand very soon OMP is constrained by orthogonality to pick seemingly irrelevant words !

	OMP	OMP with pivoting
1	and	and
2	chase	the
3	projects	for
4	ooh	chase
5	business	locum
6	own	projects
7	london	project
8	management	analysis
9	analysis	business
10	paye	technical

TABLE 3.2. First 10 terms selected by OMP and OMP with pivoting.

4 Conclusion

As a conclusion we devised a method to accelerate greatly the orthogonal matching pursuit procedure, turning it into an efficient procedure even for ultra high dimensional models. Since the resulting procedure is blazing fast, we can compensate for its relative lack of prediction efficiency compared to more involved method like the penalization methods, by using it as a preliminary estimator in a multiple stage procedure like the adaptive lasso [104] or the multi-step adaptive elastic-net [98], in the spirit of [46].

In the future, we may look at accelerating other greedy procedures relying at each step on some costly scoring step (obviously the method described here covers the case of matching pursuit too). Furthermore since the method is resilient even to a huge number of covariates it may be interesting to study its use with non linear predictors, by incorporating interaction terms in our design.

Chapter 4

A simple high-order kernel for boundary correction in density estimation

This chapter is the replica of an article submitted to a scientific review. It can be read independently from the rest of the manuscript.

Abstract

The kernel estimation method in density estimation, even if very commonly used, suffers from important difficulties as soon as the density of interest is supported on a strict sub-interval of $\mathbb{R}$, this is the so-called boundary effect. We provide here a construction of boundary kernels of any order free from the drawbacks of the boundary effect. This construction adds to some initial even kernel of a given order an odd function (which depends heavily on the initially chosen kernel) such that the order is preserved and the boundary bias is removed.

Contents

1 Introduction

1.1 Aims and Motivations

Estimating a probability density function f from an i.i.d sample $X_1, \dots, X_n$ has been an extensively studied problem (see for example [86], [91]). Despite being commonly used, the kernel density estimator, introduced in [82] [78], still suffers from important drawbacks. One of the most important, and the source of a lot of activity, is the so called boundary effect. If the density of interest is supported on some interval I of $\mathbb{R}$ with non-empty boundary, and if f admits non zero limits at those boundary points, then the usual kernel density estimator underestimates the true density at those points. In practice this is an often encountered situation, when for example the X_i's are proportions (and as a consequence take their values in $[0, 1]$), or when they are positively supported as in survival analysis, reliability, or waiting times for example.

Many methods exist which aim at correcting the boundary bias of kernel estimation. To name just a few, we can refer to the reflection method [84], [86], [25], or equivalently the cut and normalize method [50], the transformation methods [35], [96], [83], or the Beta and Gamma kernels methods [23], [12], [22], [21].

Here we will focus on the boundary kernel method of [75], [61], [60], [100], [101]. But those studies were usually concerned by the construction of order 2 kernels. We propose in this paper a simple construction of a boundary kernel of any order $l \geq 0$. The whole construction of those kernels is described in section 2 but relies on the idea of starting from any order l, compactly supported, even kernel K to which we add at any point of the boundary an odd function, such that it verifies moment conditions.

1.2 Model and Assumptions

Let $X_1, \dots, X_n$ be an i.i.d sample of n observations from a distribution $\mathbb{P}_X$ on $\mathbb{R}$, which admits a density f relative to the Lebesgue measure $\mathrm{d}x$ on $\mathbb{R}$. Let X be a random variable distributed as $\mathbb{P}_X$. Furthermore we suppose f to be compactly supported on $[-1, 1]$.

Consider K a kernel, i.e an integrable function on $\mathbb{R}$ such that $\int_{\mathbb{R}} K(u)du = 1$. For any $h > 0$ we denote by K_h the function:

$$K_h(x) = \frac{1}{h} K\Big(\frac{x}{h}\Big), \forall x \in \mathbb{R}.$$

We will always assume for simplicity that K is compactly supported on $[-1, 1]$, so that K_h is supported on $[-h, h]$. We recall the definition of the kernel estimator of f, with kernel K and bandwidth $h > 0$, defined for all $x \in \mathbb{R}$ as:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h\Big(x - X_i\Big). \tag{4.1}$$

If $h > 0$ is fixed, we can split $]-1,1[$ into disjoint subsets, $]-1,1[= B_h^l \cup I_h \cup B_h^r$, where $I_h = \{x \in]-1,1[;\ -1+h \leq x \leq 1-h\}$ will be referred to as the set of interior points, and $B_h^r = \{x \in]-1,1[;\ 1-h < x < 1\}$ is the right boundary (B_h^l is defined in a similar manner for the left boundary). We will always assume that h is small enough for the boundaries to be separated, i.e. $B_h^r \cap B_h^l = \emptyset$. This is the case as soon as $h < 1$.

1.3 Behaviour of the bias of the kernel estimator

By design, the expectation of the kernel estimator $\hat{f}_n(x)$ is the convolution product of the true density f with the kernel K_h. Using the fact that both f and K are supported on $[-1,1]$ we can write:

$$\begin{aligned}\mathbb{E}[\hat{f}_n(x)] &= \int_{\mathbb{R}} K_h(t) f(x-t) \mathrm{d}t, \\ &= \int_{\frac{x-1}{h} \vee -1}^{\frac{x+1}{h} \wedge 1} K(t) f(x-th) \mathrm{d}t.\end{aligned}$$

For any $k \geq 0$, any $h > 0$ and $x \in]-1,1[$, we define, when this quantity exists:

$$t_k(x,h) = \int_{\frac{x-1}{h} \vee -1}^{\frac{x+1}{h} \wedge 1} t^k K(t) \mathrm{d}t. \tag{4.2}$$

Given a density f supported on $[-1,1]$, we study the mean squared error (MSE) of $\hat{f}_n$ at a point $x_0 \in]-1,1[$, which is defined as:

$$\begin{aligned}\mathrm{MSE}(x_0) &= \mathbb{E}\Big[(\hat{f}_n(x_0) - f(x_0))^2\Big], \\ &= \underbrace{\Big(\mathbb{E}[\hat{f}_n(x_0)] - f(x_0)\Big)^2}_{b^2(x_0)} + \underbrace{\mathbb{E}\Big[(\hat{f}_n(x_0) - \mathbb{E}[\hat{f}_n(x_0)])^2\Big]}_{\sigma^2(x_0)}.\end{aligned}$$

Then $b^2(x_0)$ is a squared bias of estimation at x_0, while $\sigma^2(x_0)$ is the variance of the estimator at x_0. It is well known that the variance of a kernel estimator can be bounded with minimal hypotheses on f. Precisely as soon as f is bounded and the kernel K is square integrable, we can show that for any $x_0 \in \mathbb{R}$, for any $h > 0$ and $n \geq 1$:

$$\sigma^2(x_0) \leq \frac{C}{nh}, \tag{4.3}$$

for some constant $C > 0$ depending on $\|f\|_\infty$ and the kernel.

The main question is then to understand the bias term in the MSE. To have precise bounds on the bias we need an extra assumption on the smoothness of the density f. We will work here under the hypothesis of Hölder smoothness of the density restricted to $]-1,1[$. Precisely let β and L be two positive numbers. We consider the Hölder class $\Sigma(\beta, L,]-1,1[)$ on $]-1,1[$, which is the set of $l = \lfloor \beta \rfloor$ times differentiable functions

$g :]-1,1[\to \mathbb{R}$, with all derivatives of order k, $0 \leq k \leq l$, are bounded and whose derivative $g^{(l)}$ satisfies:

$$|f^{(l)}(y) - f^{(l)}(x)| \leq L|y-x|^{\beta-l}, \forall x, y \in]-1,1[. \tag{4.4}$$

We describe the boundary bias phenomenon in the next lemma, for Hölder smooth densities.

Lemma 4.1. *Suppose that $f \in \Sigma(\beta, L,]-1,1[)$, then there exists τ, $|\tau| \leq 1$, such that for all $x_0 \in [-1,1]$:*

$$b(x_0) = f(x_0)(t_0(x_0,h)-1) + \sum_{k=1}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0) t_k(x_0,h) h^k + \frac{(-1)^l}{l!} h^l \int_{\frac{x_0-1}{h} \vee -1}^{\frac{x_0+1}{h} \wedge 1} t^l K(t) f^{(l)}(x_0 - \tau t h) \mathrm{d}t,$$

where $l = \lfloor \beta \rfloor$.

For a proof we refer to section 5.1.

High-order kernels are a construction which allows to cancel the intermediate terms in the expansion of lemma 4.1, at least when the density f is globally Hölder smoooth on the whole real line. Indeed we say that K is a kernel of order l if $\int_{\mathbb{R}} u^k K(u) \mathrm{d}u = 0$, for all $1 \leq k \leq l$. From now on we will suppose that K is an order l kernel supported on $[-1,1]$ (for such a construction we refer to [91]). In the case where f is supported on $[-1,1]$ and given such a kernel, lemma 4.1 has the following implications:

— if $x_0 \in I_h$, i.e. if x_0 is an interior point for h, then:

$$t_k(x,h) = \int_{\frac{x-1}{h} \vee -1}^{\frac{x+1}{h} \wedge 1} t^k K(t) \mathrm{d}t = \int_{-1}^{1} t^k K(t) \mathrm{d}t = 0,$$

for all $1 \leq k \leq l$, since K is an order l kernel. This implies that on interior points the estimator $\hat{f}_n$ will behave as a kernel estimator of a density with no boundary.

— now consider $x_0 \in B_h^r$, the right boundary (the same reasoning applies obviously to the left boundary). There exists $\alpha \in [0,1]$ such that $x_0 = x_\alpha = 1 - \alpha h$, and

$$t_k(x_\alpha, h) = \int_{-\alpha}^{1} u^k K(u) \mathrm{d}u.$$

As a consequence in the right boundary the bias behaves differently than on the interior point. Even worse the kernel estimator may not even be consistent when the bandwidth goes to 0. For example in the case where $K(u) = \frac{1}{2} 1\{-1 \leq u \leq 1\}$, $t_0(x_{1/2}, h) = 3/4$, so that the kernel estimator is biased with a bias $|b(x_{1/2})| \to \frac{f(1^+)}{4}$ as $h \to 0$.

The following section describes our construction of a boundary kernel free from those downsides.

2 Boundary kernel modification

2.1 Folding

Following the construction from [91] we consider an even kernel K, supported on $[-1,1]$, of order $2l+1$, where $l \geq 0$ is an integer. As shown in the previous section those kernels are well adapted to densities Hölder smooths on the whole real line, or, when the density is compactly supported, to estimate it on interior points (relatively to the chosen bandwidth h). We consider now a modification of K which allows it to have the correct bias on boundary points for densities in $\Sigma(\beta, L,]-1,1[)$. We will focus on the right boundary but obviously our reasoning extends to the left one.

Let $\alpha \in]0,1]$, and let $x_\alpha = 1 - \alpha h$ be a point of the right boundary B_h^r. We will construct the modification K_α of K, adapted to the estimation of a density at the point x_α, in a simple additive form. Indeed we will look for a function $\tilde{K}_\alpha$ such that:

$$K_\alpha(u) = K(u) + C_\alpha \tilde{K}_\alpha(u), \tag{4.5}$$

where $\tilde{K}_\alpha$ is supported on $[-1,1]$ and is an odd function (so that we can restrict its description to $[0,1]$), and C_α is a normalizing constant such that $C_\alpha \int_0^1 \tilde{K}_\alpha(u)\mathrm{d}u = 1/2$.

The description of K_α will be done in two steps. First we define it on $[\alpha, 1]$:

$$\tilde{K}_\alpha(u) = K(u), \ \forall u \in [\alpha, 1]. \tag{4.6}$$

Then for any $0 \leq k \leq l$, the even boundary moments verify:

$$t_{2k}(x_\alpha, h) = \int_{-\alpha}^1 u^k K_\alpha(u)\mathrm{d}u = \int_{-\alpha}^1 u^k K(u)\mathrm{d}u + \int_{-\alpha}^1 u^k \tilde{K}_\alpha(u)\mathrm{d}u = \int_{-1}^1 u^k K(u)\mathrm{d}u.$$

Indeed since K is even, $u \mapsto u^{2k}K(u)$ is even, and $u \mapsto u^{2k}\tilde{K}_\alpha(u)$ is odd. This implies immediatly by eq. (4.6) that:

$$\int_{-\alpha}^1 u^k \tilde{K}_\alpha(u)\mathrm{d}u = \int_{\alpha}^1 u^{2k} \tilde{K}_\alpha(u)\mathrm{d}u = \int_{-1}^{-\alpha} u^{2k} K(u)\mathrm{d}u. \tag{4.7}$$

This proves that the condition eq. (4.6) is enough to ensure that K_α has all its even boundary moments in x_α equal to those of K in the interior points, as long as $\tilde{K}_\alpha$ is odd, whatever the values it takes on $[0, \alpha[$. In particular it is enough to take $\tilde{K}_\alpha(u) = 0$ for $u \in [0, \alpha[$ to ensure consistent estimation in the boundary. This is the reflexion trick, and it extends in fact to all even moments.

We still have to define $\tilde{K}_\alpha$ on $[0, \alpha[$ in such a manner that the odd boundary moments will be 0. To do so we will look for a polynomial function of order l in $[0, \alpha[$, i.e. we will look for $\tilde{K}_\alpha$ of the form:

$$\tilde{K}_\alpha(u) = \sum_{j=0}^{l} a_j u^j, \ \forall u \in [0, \alpha[. \tag{4.8}$$

Then for any $0 \leq k \leq l$, we can write the odd boundary moments as:

$$t_{2k+1}(x_\alpha, h) = \int_{-\alpha}^{1} u^{2k+1} K(u) \mathrm{d}u + \int_{-\alpha}^{1} u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u. \tag{4.9}$$

Now since $u \mapsto u^{2k+1} K(u)$ is odd, and $u \mapsto u^{2k+1} \tilde{K}_\alpha(u)$ is even for any $0 \leq k \leq l$, we can define:

$$m_k(\alpha) = \int_{-\alpha}^{1} u^{2k+1} K(u) \mathrm{d}u = \int_{\alpha}^{1} u^{2k+1} K(u) \mathrm{d}u, \tag{4.10}$$

and we can notice that from eq. (4.6) the odd boundary moments of $\tilde{K}_\alpha$ verify:

$$\int_{-\alpha}^{1} u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u = m_k(\alpha) + 2 \int_{0}^{\alpha} u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u. \tag{4.11}$$

Then from eq. (4.9) we can deduce the following linear relation:

$$t_{2k+1}(x_\alpha, h) = 2m_k(\alpha) + 2 \int_{0}^{\alpha} u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u. \tag{4.12}$$

Finally using eq. (4.8) we have that:

$$\int_{0}^{\alpha} u^{2k+1} \tilde{K}_\alpha(u) \mathrm{d}u = \sum_{j=0}^{l} \alpha_j \int_{0}^{\alpha} u^{2k+j+1} \mathrm{d}u = \sum_{j=0}^{l} a_j \frac{\alpha^{2k+j+2}}{2k+j+1}. \tag{4.13}$$

Since we want that $t_{2k+1}(x_\alpha, h) = 0$ for any $0 \leq k \leq l$, we want the coefficients a_j to verify:

$$\sum_{j=0}^{l} a_j \frac{\alpha^{2k+j+2}}{2k+j+1} = -m_k(\alpha), \ \forall\, 0 \leq k \leq l. \tag{4.14}$$

This system of equations can be rewritten in matrix form, where the coefficients a_j form a vector $\mathbf{a}$, the quantities $m_k(\alpha)$ a vector $\mathbf{m}(\alpha)$ related by:

$$\Lambda(\alpha)\mathbf{a} = \mathbf{m}(\alpha), \tag{4.15}$$

where $\Lambda(\alpha)$ is a $(l+1) \times (l+1)$ matrix with $\Lambda_{kj}(\alpha) = \frac{\alpha^{2k+j+2}}{2k+j+1}$.

Lemma 4.2. *The matrix $\Lambda(\alpha)$ is invertible for all $\alpha \in (0, 1]$.*

For a proof we refer to section 5.2. Even if this lemma implies the existence of our construction for any $\alpha \in (0, 1]$, the resolution of eq. (4.17) is very instable numerically. The nex subsection introduces a slight modification which stabilizes the resolution step, by developping $\tilde{K}_\alpha$ on an orthogonal family of polynomials.

2.2 Expansion of the solution on an orthogonal basis

In order to solve a better conditioned system than the one we obtained with looking at $\tilde{K}_\alpha$ as a sum of monomials, we will look for $\tilde{K}_\alpha$ as a sum of orthogonal polynomials. Let the L_k's be Legendre polynomials ([87]), which form an orthogonal family in $L^2([-1, 1])$ (we choose to normalize them so that they have unit L^2 norm too). Then the functions:

$$\tilde{L}_k : x \mapsto \frac{1}{\sqrt{\alpha}} L_k\Big(\frac{x}{\alpha}\Big), \ \forall\, x \in [-\alpha, \alpha],$$

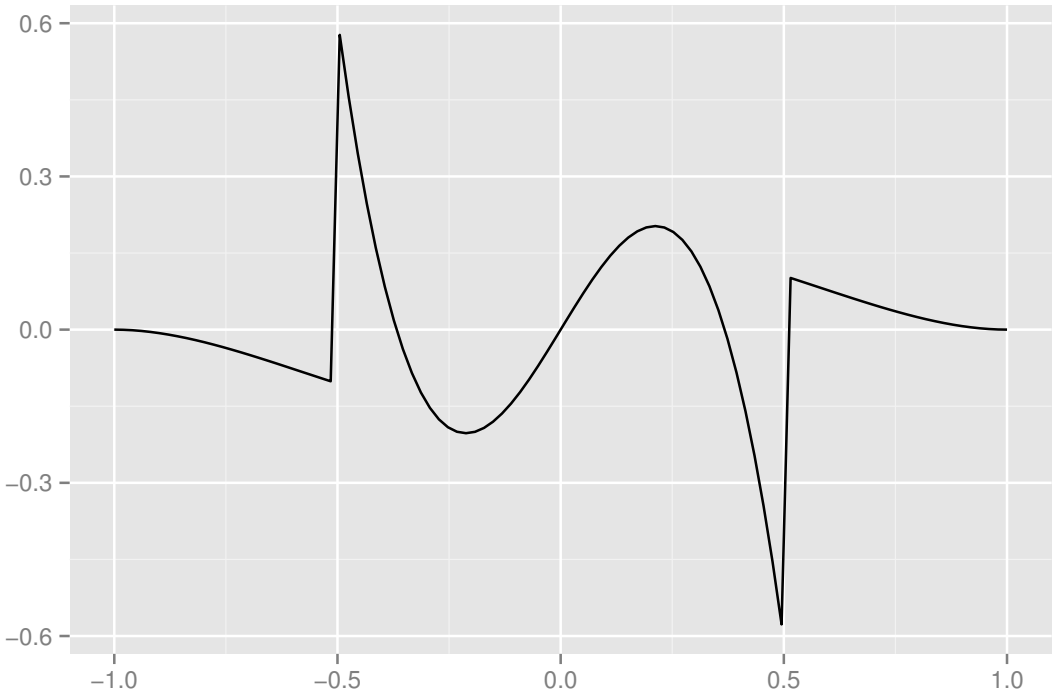

FIGURE 4.1. Correction $\tilde{K}_\alpha$ for an order 3 kernel at $\alpha = 0.5$.

are an orthogonal unitary family in $L^2([-\alpha, \alpha])$.

Now let us look for $\tilde{K}_\alpha$ as a sum of those orthogonal polynomials:

$$\tilde{K}_\alpha(u) = \sum_{i=0}^{l} b_i \tilde{L}_{2i+1}(u), \ \forall\, u \in [-\alpha, \alpha]. \tag{4.16}$$

Then using the decomposition 4.16 of $\tilde{K}_\alpha$, we obtain again the coefficients b_i as solutions to the linear system:

$$\mathbf{L}(\alpha)\mathbf{b} = \mathbf{m}(\alpha), \tag{4.17}$$

where the matrix $L(\alpha)$ coefficients are $\Lambda_{ij}(\alpha) = \alpha^{2i+1/2} \int_0^1 L_{2j+1}(y) y^{2i+1} \mathrm{d}y$. This resolution behaves much better than 4.17, and allows us to build efficiently our boundary kernels. For example fig. 4.1 and fig. 4.2 illustrate, respectively, the modification $\tilde{K}_{1/2}$ and the final kernel $K_{1/2}$, starting from an order 3 kernel.

3 Numerical Study

To illustrate the empirical efficiency of our method, we propose to estimate the density f_X of a truncated, between -1 and 1, normal random variable, with mean 0 and standard deviation 4. To do so we fix the bandwidth (here $h = 0.6$) and focus on the right boundary (here the interval $[0.4, 1]$). We then compare an Epanechnikov kernel estimator (i.e. using the order 1 kernel $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \leq 1\}$, very common in practice) and our

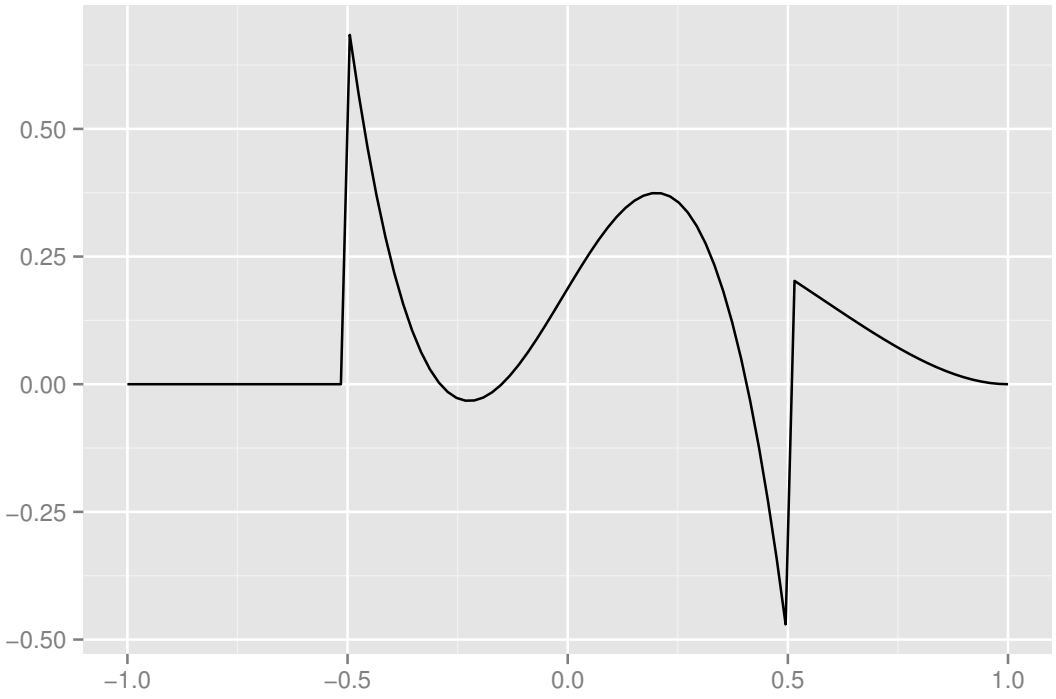

FIGURE 4.2. Modified kernel K_α for an intial order 3 kernel at $\alpha = 0.5$.

methodology. The figure 4.3 reports such a comparison.

To provide a better measurement of the quality of our procedure we proceed in the following way: we repeat $K = 100$ times the following experiment

1. we draw 1000 observations of density f_X (idependantly),
2. we compute the kernel estimator $\hat{f}_h$ and its modification with our procedure $\tilde{f}_h$ with a common h (here $h = 0.6$),
3. we compute the quantities $P_{\hat{f}} = \int_{1-h}^{1} \hat{f}_h(x)\mathrm{d}x$ and $P_{\tilde{f}} = \int_{1-h}^{1} \tilde{f}_h(x)\mathrm{d}x$, which are the estimated probabilities for the right boundary by the two methods.

We compare those estimated probabilities to the true probability $P = \int_{1-h}^{1} f_X(x)\mathrm{d}x$ of the right boundary, i.e. we compute $\frac{P-P_{\hat{f}}}{P} \times 100$ and $\frac{P-P_{\tilde{f}}}{P} \times 100$, and report the results (in %) on 4.1.

	$\frac{P-P_{\hat{f}}}{P} \times 100$	$\frac{P-P_{\tilde{f}}}{P} \times 100$
Mean	-18.40 %	7.63 %
Standard Deviation	2.89 %	3.70 %

TABLE 4.1. Summary of 100 repetitions of estimating the right boundary of a truncated normal random variable with fix $h = 0.6$.

We can see that our procedure greatly improves the usual kernel estimation. Indeed kernel estimation tends to underestimate the right boundary probability, as we already explained, while our procedure tends to slightly overestimate it, but very less so. This

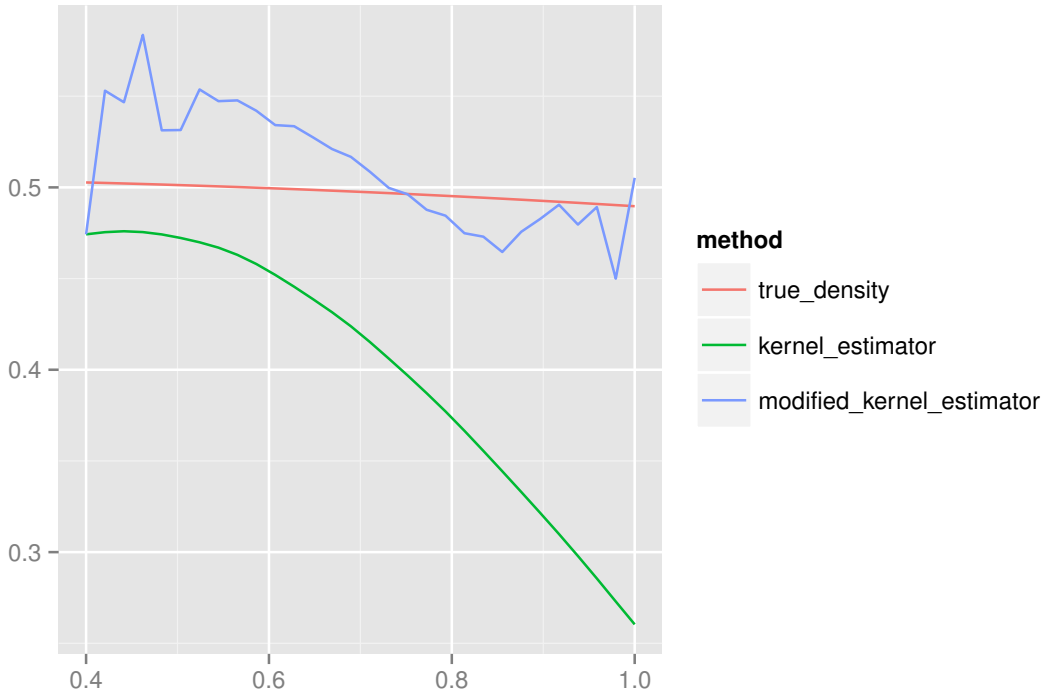

FIGURE 4.3. Density estimation of a truncated normal random variable. We have 1000 observations X_i, and compare a kernel estimator with its modification at $h = 0.6$ for the right boundary.

improvement is paid with a slight increase of the variance (since we increase the L^2 norm of the iniital kernel this had to be expected), but clearly the gains in mean outperform this new instability.

4 Conclusion

We described a general methodology to modify any kernel, in order to adapt it to the boundary of the true density to be estimated, while keeping its initial order. This procedure performs uniformly better than usual kernel estiamtion in the boundary, with fixed bandwidth. Then if we use an adaptive (but global) bandwidth selection procedure (like [65] or cross-validation), we can restrict ourselves to use only the interior points common to all the bandwith in a grid. It remains to study how to adapt the methods which use local bandwiths to this situation.

5 Proofs

5.1 Proof of lemma 4.1

We have already noticed that for any $x_0 \in]-1, 1[$:

$$\mathbb{E}[\hat{f}_n(x_0)] = \int_{\frac{x_0-1}{h}\vee -1}^{\frac{x_0+1}{h}\wedge 1} K(t) f(x_0 - th) \mathrm{d}t.$$

Since $f \in \Sigma(\beta, L,]-1, 1[)$, we have by Taylor expansion around x_0, that there exists τ, $|\tau| \leq 1$, such that for all $\frac{x_0-1}{h} \leq t \leq \frac{x_0+1}{h}$:

$$f(x_0 - th) = \sum_{k=0}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0)(th)^k + \frac{(-1)^l}{l!}(th)^l f^{(l)}(x_0 - \tau th).$$

Consequently:

$$\begin{aligned}
\mathbb{E}[\hat{f}_n(x_0)] &= \sum_{k=0}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0) t_k(x_0, h) h^k + \frac{(-1)^l}{l!} h^l \int_{\frac{x_0-1}{h}\vee -1}^{\frac{x_0+1}{h}\wedge 1} t^l K(t) f^{(l)}(x_0 - \tau th) \mathrm{d}t, \\
&= f(x_0) t_0(x_0, h) + \sum_{k=1}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0) t_k(x_0, h) h^k + \frac{(-1)^l}{l!} h^l \int_{\frac{x_0-1}{h}\vee -1}^{\frac{x_0+1}{h}\wedge 1} t^l K(t) f^{(l)}(x_0 - \tau th) \mathrm{d}t.
\end{aligned}$$

Finally :

$$b(x_0) = f(x_0)(t_0(x_0, h) - 1) + \sum_{k=1}^{l-1} \frac{(-1)^k}{k!} f^{(k)}(x_0) t_k(x_0, h) h^k + \frac{(-1)^l}{l!} h^l \int_{\frac{x_0-1}{h}\vee -1}^{\frac{x_0+1}{h}\wedge 1} t^l K(t) f^{(l)}(x_0 - \tau th) \mathrm{d}t.$$

5.2 Proof of lemma 4.2

Let, for all $\alpha \in (0, 1]$, $P(\alpha) = \det(\Lambda(\alpha))$, which is a polynomial in α. Furthermore, Leibniz formula immediatly implies that:

$$P(\alpha) = C\alpha^{\frac{3l(l+1)}{2}+2}, \tag{4.18}$$

for some constant $C = P(1)$, and for all $\alpha \in (0, 1]$. It remains to prove that $C \neq 0$, which implies that the matrix $\Lambda(\alpha)$ is invertible for any non zero α. But $\Lambda(1)$ is a Cauchy matrix, with general term $\Lambda(1)_{kj} = \frac{1}{2k+j+1}$, and consequently is invertible.

Bibliography

[1] Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone, *Adapting to unknown sparsity by controlling the false discovery rate*, Ann. Statist. **34** (2006), no. 2, 584–653. MR 2281879 (2008c:62012)

[2] Afonso S. Bandeira, Matthew Fickus, Dustin G. Mixon, and Percy Wong, *The road to deterministic matrices with the restricted isometry property*, J. Fourier Anal. Appl. **19** (2013), no. 6, 1123–1149. MR 3132908

[3] Richard Beals and Roderick Wong, *Special functions*, Cambridge Studies in Advanced Mathematics, vol. 126, Cambridge University Press, Cambridge, 2010, A graduate text. MR 2683157 (2011j:33001)

[4] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen, *Sparse models and methods for optimal instruments with an application to eminent domain*, Econometrica **80** (2012), no. 6, 2369–2429.

[5] Karine Bertin and Nicolas Klutchnikoff, *Minimax properties of beta kernel estimators*, Journal of Statistical Planning and Inference **141** (2011), no. 7, 2287 – 2297.

[6] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov, *Simultaneous analysis of lasso and Dantzig selector*, Ann. Statist. **37** (2009), no. 4, 1705–1732. MR 2533469 (2010j:62118)

[7] Åke Björck, *Numerical methods for least squares problems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. MR 1386889 (97g:65004)

[8] ———, *Numerical methods in matrix computations*, Texts in Applied Mathematics, vol. 59, Springer, Cham, 2015. MR 3288840

[9] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, *Kernel density estimation via diffusion*, Ann. Statist. **38** (2010), no. 5, 2916–2957. MR 2722460 (2011k:62098)

[10] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities*, Oxford University Press, Oxford, 2013, A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR 3185193

[11] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, New York, NY, USA, 2004.

[12] Bruce M. Brown and Song Xi Chen, *Beta-Bernstein smoothing for regression curves with compact support*, Scand. J. Statist. **26** (1999), no. 1, 47–59. MR 1685301 (2000a:62097)

[13] Peter Bühlmann and Sara van de Geer, *Statistics for high-dimensional data*, Springer Series in Statistics, Springer, Heidelberg, 2011, Methods, theory and applications. MR 2807761 (2012e:62006)

[14] Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electron. J. Stat. **1** (2007), 169–194. MR 2312149 (2008h:62101)

[15] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp, *Sparse density estimation with ℓ_1 penalties*, Learning theory, Lecture Notes in Comput. Sci., vol. 4539, Springer, Berlin, 2007, pp. 530–543. MR 2397610 (2009f:68085)

[16] T.T. Cai and Lie Wang, *Orthogonal matching pursuit for sparse signal recovery with noise*, Information Theory, IEEE Transactions on **57** (2011), no. 7, 4680–4688.

[17] Emmanuel Candes and Terence Tao, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35** (2007), no. 6, 2313–2351. MR 2382644 (2009b:62016)

[18] Emmanuel J. Candès, *Modern statistical estimation via oracle inequalities*, Acta Numer. **15** (2006), 257–325. MR 2269743 (2007m:62095)

[19] Emmanuel J. Candes and Terence Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215. MR 2243152 (2007b:94313)

[20] ———, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215. MR 2243152 (2007b:94313)

[21] Song Xi Chen, *Beta kernel estimators for density functions*, Computational Statistics and Data Analysis **31** (1999), no. 2, 131 – 145.

[22] Song Xi Chen, *Beta kernel smoothers for regression curves*, Statist. Sinica **10** (2000), no. 1, 73–91. MR 1742101 (2000k:62071)

[23] ———, *Probability density function estimation using gamma kernels*, Ann. Inst. Statist. Math. **52** (2000), no. 3, 471–480. MR 1794247 (2001h:62061)

[24] Ming-Yen Cheng, Jianqing Fan, and J. S. Marron, *On automatic boundary corrections*, Ann. Statist. **25** (1997), no. 4, 1691–1708. MR 1463570 (98k:62049)

[25] Daren B. H. Cline and Jeffrey D. Hart, *Kernel estimation of densities with discontinuities or discontinuous derivatives*, Statistics **22** (1991), no. 1, 69–84. MR 1097362 (92e:62069)

[26] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), no. 1, 211–231. MR 2449058 (2010d:94024)

[27] Ann Cowling and Peter Hall, *On pseudodata methods for removing boundary effects in kernel density estimation*, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 3, 551–563. MR 1394366 (97a:62076)

[28] Ingrid Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992. MR 1162107 (93e:42045)

[29] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Constr. Approx. **13** (1997), no. 1, 57–98. MR 1424364 (98a:65014)

[30] Geoffrey M. Davis, Stephane G. Mallat, and Zhifeng Zhang, *Adaptive time-frequency decompositions*, Optical Engineering **33** (1994), no. 7, 2183–2191.

[31] Holger Dette and Jens Wagener, *Least squares estimation in high dimensional sparse heteroscedastic models*, Robustness and complex data structures, Springer, Heidelberg, 2013, pp. 135–147. MR 3135878

[32] R. A. DeVore and V. N. Temlyakov, *Nonlinear approximation in finite-dimensional spaces*, J. Complexity **13** (1997), no. 4, 489–508. MR 1606541 (99c:41053)

[33] Ronald A. DeVore, *Nonlinear approximation*, Acta numerica, 1998, Acta Numer., vol. 7, Cambridge Univ. Press, Cambridge, 1998, pp. 51–150. MR 1689432 (2001a:41034)

[34] Luc Devroye, *A course in density estimation*, Progress in Probability and Statistics, vol. 14, Birkhäuser Boston, Inc., Boston, MA, 1987. MR 891874 (88d:62070)

[35] Luc Devroye and László Györfi, *Nonparametric density estimation*, Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics, John Wiley & Sons, Inc., New York, 1985, The L_1 view. MR 780746 (86i:62065)

[36] David L. Donoho and Xiaoming Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Trans. Inform. Theory **47** (2001), no. 7, 2845–2862. MR 1872845 (2002k:94012)

[37] David L. Donoho and Iain M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), no. 3, 425–455. MR 1311089 (95m:62076)

[38] David L. Donoho and Iain M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[39] David L. Donoho and Iain M. Johnstone, *Neo-classical minimax problems, thresholding and adaptive function estimation*, Bernoulli **2** (1996), no. 1, 39–62. MR 1394051 (97f:62062)

[40] ———, *Minimax estimation via wavelet shrinkage*, Ann. Statist. **26** (1998), no. 3, 879–921. MR 1635414 (99i:62086)

[41] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard, *Wavelet shrinkage: asymptopia?*, J. Roy. Statist. Soc. Ser. B **57** (1995), no. 2, 301–369, With discussion and a reply by the authors. MR 1323344 (96g:62068)

[42] David L. Donoho, Iain M. Johnstone, Gerard Kerkyacharian, and Dominique Picard, *Wavelet shrinkage: asymptopia*, Journal of the Royal Statistical Society, Ser. B (1995), 371–394.

[43] D.L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, *Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit*, Information Theory, IEEE Transactions on **58** (2012), no. 2, 1094–1121.

[44] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), no. 2, 407–499, With discussion, and a rejoinder by the authors. MR 2060166 (2005d:62116)

[45] Jianqing Fan and Irène Gijbels, *Variable bandwidth and local linear regression smoothers*, Ann. Statist. **20** (1992), no. 4, 2008–2036. MR 1193323 (94a:62050)

[46] Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space*, J. R. Stat. Soc. Ser. B Stat. Methodol. **70** (2008), no. 5, 849–911. MR 2530322

[47] Dean P. Foster and Edward I. George, *The risk inflation criterion for multiple regression*, Ann. Statist. **22** (1994), no. 4, 1947–1975. MR 1329177 (96c:62119)

[48] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, 2009.

[49] T. Gasser, H-G. Muller, and V. Mammitzsch, *Kernels for nonparametric curve estimation*, Journal of the Royal Statistical Society. Series B (Methodological) **47** (1985), no. 2, pp. 238–252 (English).

[50] Theo Gasser and Hans-Georg Müller, *Kernel estimation of regression functions*, Smoothing Techniques for Curve Estimation (Th. Gasser and M. Rosenblatt, eds.), Lecture Notes in Mathematics, vol. 757, Springer Berlin Heidelberg, 1979, pp. 23–68 (English).

[51] Gery Geenens, *Probit transformation for kernel density estimation on the unit interval*, J. Amer. Statist. Assoc. **109** (2014), no. 505, 346–358. MR 3180568

[52] Anna C. Gilbert, S. Muthukrishnan, and Martin J. Strauss, *Approximation of functions over redundant dictionaries using coherence*, Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (Baltimore, MD, 2003), ACM, New York, 2003, pp. 243–252. MR 1974925

[53] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985. MR 832183 (87e:15001)

[54] Jinzhu Jia, Karl Rohe, and Bin Yu, *The lasso under heteroscedasticity*, arXiv:1011.1026 [stat] (2010), arXiv: 1011.1026.

[55] Iain M. Johnstone, *Minimax Bayes, asymptotic minimax and sparse wavelet priors*, Statistical decision theory and related topics, V (West Lafayette, IN, 1992), Springer, New York, 1994, pp. 303–326. MR 1286310 (95d:62012)

[56] ———, *Wavelet shrinkage for correlated data and inverse problems: adaptivity results*, Statist. Sinica **9** (1999), no. 1, 51–83. MR 1678881 (2000g:62087)

[57] Iain M. Johnstone, *Wavelet shrinkage for correlated data and inverse problems: adaptivity results*, Statist. Sinica (1999), 51–83.

[58] Iain M. Johnstone and Bernard W. Silverman, *Wavelet threshold estimators for data with correlated noise*, 1994.

[59] Iain M. Johnstone and Bernard W. Silverman, *Wavelet threshold estimators for data with correlated noise*, J. Roy. Statist. Soc. Ser. B **59** (1997), no. 2, 319–351. MR 1440585 (98h:62054)

[60] M. C. Jones and P. J. Foster, *A simple nonnegative boundary correction method for kernel density estimation*, Statist. Sinica **6** (1996), no. 4, 1005–1013. MR 1422417 (97k:62089)

[61] M.C. Jones, *Simple boundary correction for kernel density estimation*, Statistics and Computing **3** (1993), no. 3, 135–146 (English).

[62] Yitzhak Katznelson, *An introduction to harmonic analysis*, third ed., Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2004. MR 2039503 (2005d:43001)

[63] Gérard Kerkyacharian, Mathilde Mougeot, Dominique Picard, and Karine Tribouley, *Learning out of leaders*, Multiscale, nonlinear and adaptive approximation, Springer, Berlin, 2009, pp. 295–324. MR 2648377 (2011k:62114)

[64] Gérard Kerkyacharian and Dominique Picard, *Thresholding algorithms, maxisets and well-concentrated bases*, Test **9** (2000), no. 2, 283–344, With comments, and a rejoinder by the authors. MR 1821645 (2002b:62043)

[65] O. V. Lepskii, *On a problem of adaptive estimation in gaussian white noise*, Theory of Probability & Its Applications **35** (1991), no. 3, 454–466.

[66] Entao Liu and Vladimir N. Temlyakov, *The orthogonal super greedy algorithm and applications in compressed sensing*, IEEE Trans. Inform. Theory **58** (2012), no. 4, 2040–2047. MR 2951314

[67] ———, *Super greedy type algorithms*, Adv. Comput. Math. **37** (2012), no. 4, 493–504. MR 2988776

[68] S.G. Mallat and Zhifeng Zhang, *Matching pursuits with time-frequency dictionaries*, Trans. Sig. Proc. **41** (1993), no. 12, 3397–3415.

[69] Pascal Massart, *Concentration inequalities and model selection*, Lecture Notes in Mathematics, vol. 1896, Springer, Berlin, 2007, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. MR 2319879 (2010a:62008)

[70] Yves Meyer, *Ondelettes et opérateurs. I*, Actualités Mathématiques. [Current Mathematical Topics], Hermann, Paris, 1990, Ondelettes. [Wavelets]. MR 1085487 (93i:42002)

[71] Alan Miller, *Subset selection in regression*, second ed., Monographs on Statistics and Applied Probability, vol. 95, Chapman & Hall/CRC, Boca Raton, FL, 2002. MR 2001193

[72] M. Mougeot, D. Picard, and K. Tribouley, *LOL selection in high dimension*, Comput. Statist. Data Anal. **71** (2014), 743–757. MR 3132003

[73] Mathilde Mougeot, Dominique Picard, and Karine Tribouley, *Learning out of leaders*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **74** (2012), no. 3, 475–513. MR 2925371

[74] ———, *Grouping strategies and thresholding for high dimensional linear models*, J. Statist. Plann. Inference **143** (2013), no. 9, 1417–1438. MR 3070245

[75] Hans-Georg Müller, *Smooth optimum kernel estimators near endpoints*, Biometrika **78** (1991), no. 3, 521–530. MR 1130920 (92f:62052)

[76] D. Needell and J. A. Tropp, *CoSaMP: iterative signal recovery from incomplete and inaccurate samples*, Appl. Comput. Harmon. Anal. **26** (2009), no. 3, 301–321. MR 2502366 (2010c:94018)

[77] Deanna Needell and Roman Vershynin, *Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit*, Found. Comput. Math. **9** (2009), no. 3, 317–334. MR 2496554 (2009m:65270)

[78] Emanuel Parzen, *On estimation of a probability density function and mode*, Ann. Math. Statist. **33** (1962), 1065–1076. MR 0143282 (26 #841)

[79] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, *Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition*, Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on, Nov 1993, pp. 40–44 vol.1.

[80] Gilles Pisier, *Some applications of the metric entropy condition to harmonic analysis*, Banach spaces, harmonic analysis, and probability theory (Storrs, Conn., 1980/1981), Lecture Notes in Math., vol. 995, Springer, Berlin, 1983, pp. 123–154. MR 717231 (85f:60061)

[81] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Trans. Inform. Theory **57** (2011), no. 10, 6976–6994. MR 2882274 (2012k:62204)

[82] Murray Rosenblatt, *Remarks on some nonparametric estimates of a density function*, Ann. Math. Statist. **27** (1956), 832–837. MR 0079873 (18,159f)

[83] David Ruppert and Daren B. H. Cline, *Bias reduction in kernel density estimation by smoothed empirical transformations*, Ann. Statist. **22** (1994), no. 1, 185–210. MR 1272080 (95g:62078)

[84] Eugene F. Schuster, *Incorporating support constraints into nonparametric estimators of densities*, Communications in Statistics - Theory and Methods **14** (1985), no. 5, 1123–1136.

[85] George A. F. Seber and Alan J. Lee, *Linear regression analysis*, second ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. MR 1958247 (2003m:62004)

[86] Bernard W. Silverman, *Density estimation for statistics and data analysis*, CRC Press, April 1986 (en).

[87] Gábor Szegő, *Orthogonal polynomials*, fourth ed., American Mathematical Society, Providence, R.I., 1975, American Mathematical Society, Colloquium Publications, Vol. XXIII. MR 0372517 (51 #8724)

[88] Vladimir Temlyakov, *Greedy approximation*, Cambridge Monographs on Applied and Computational Mathematics, vol. 20, Cambridge University Press, Cambridge, 2011. MR 2848161

[89] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1994), 267–288.

[90] Joel A. Tropp, *Greed is good: algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory **50** (2004), no. 10, 2231–2242. MR 2097044 (2005e:94036)

[91] Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer, November 2008 (en).

[92] ———, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR 2724359 (2011g:62006)

[93] Sara van de Geer, Peter Bühlmann, and Shuheng Zhou, *The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)*, Electron. J. Stat. **5** (2011), 688–749. MR 2820636 (2012i:62201)

[94] J. Wagener and H. Dette, *Bridge estimators and the adaptive Lasso under heteroscedasticity*, Math. Methods Statist. **21** (2012), no. 2, 109–126. MR 2974012

[95] ———, *The adaptive Lasso in high-dimensional sparse heteroscedastic models*, Math. Methods Statist. **22** (2013), no. 2, 137–154. MR 3071959

[96] M. P. Wand, J. S. Marron, and D. Ruppert, *Transformations in density estimation*, J. Amer. Statist. Assoc. **86** (1991), no. 414, 343–361, With discussion and a rejoinder by the authors. MR 1137118 (92i:62081)

[97] Michael J. Wichura, *The coordinate-free approach to linear models*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2006. MR 2283455 (2008h:62008)

[98] Nan Xiao and Qing-Song Xu, *Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection*, Journal of Statistical Computation and Simulation **0** (0), no. 0, 1–11.

[99] Shunpu Zhang, *A note on the performance of the gamma kernel estimators at the boundary*, Statistics and Probability Letters **80** (2010), no. 7–8, 548 – 557.

[100] Shunpu Zhang and Rohana J. Karunamuni, *On kernel density estimation near endpoints*, J. Statist. Plann. Inference **70** (1998), no. 2, 301–316. MR 1649872 (99k:62080)

[101] ———, *On nonparametric density estimation at the boundary*, J. Nonparametr. Statist. **12** (2000), no. 2, 197–221. MR 1752313 (2000m:62023)

[102] Tong Zhang, *Sparse recovery with orthogonal matching pursuit under RIP*, IEEE Trans. Inform. Theory **57** (2011), no. 9, 6215–6221. MR 2857968 (2012h:94074)

[103] Hui Zou, *The adaptive lasso and its oracle properties*, J. Amer. Statist. Assoc. **101** (2006), no. 476, 1418–1429. MR 2279469 (2008d:62024)

[104] Hui Zou, *The adaptive lasso and its oracle properties*, Journal of the American Statistical Association **101** (2006), no. 476, 1418–1429.

[105] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, Series B **67** (2005), 301–320.